

Informe Proyecto 2 entrega 1

Juan Luis Solórzano (carnet: 201598)

Micaela Yataz (carnet: 18960)

2025-01-20

git: https://github.com/JusSolo/Mineria_Proyecto2.git

Análisis exploratorio

descripción de las variables:

- **SalePrice:** cuantitativa, el precio de venta de la propiedad
- **MSSubClass:** cuantitativa, la clase de construcción
- **MSZoning:** cualitativa, la clasificación general de zona
- **LotFrontage:** cuantitativa, pies lineales de calle conectada a la propiedad
- **LotArea:** cuantitativa, tamaño del lote en pies cuadrados
- **Street:** cualitativa, tipo de acceso vial
- **Alley:** tipo de acceso por callejón
- **LotShape:** cualitativa, forma general de la propiedad
- **LandContour:** cualitativa, relieve de la propiedad
- **Utilities:** cualitativa, tipo de servicios públicos disponibles
- **LotConfig:** cualitativa, configuración del lote
- **LandSlope:** cualitativa, pendiente de la propiedad
- **Neighborhood:** cualitativa, ubicaciones físicas dentro de los límites de la ciudad de Ames
- **Condition1:** cualitativa, proximidad a la carretera principal o ferrocarril
- **Condition2:** cualitativa, proximidad a la carretera principal o ferrocarril (si hay un segundo presente)
- **BldgType:** cualitativa, tipo de vivienda
- **HouseStyle:** cualitativa, estilo de vivienda
- **OverallQual:** cuantitativa, calidad general de materiales y acabados
- **OverallCond:** cuantitativa, calificación general de la condición
- **YearBuilt:** cuantitativa, año de construcción original
- **YearRemodAdd:** cuantitativa, año de remodelación
- **RoofStyle:** cualitativa, tipo de techo
- **RoofMatl:** cualitativa, material del techo
- **Exterior1st:** cualitativa, revestimiento exterior de la casa

- Exterior2nd: cualitativa, revestimiento exterior de la casa (si se utiliza más de un material)
- MasVnrType: cualitativa, tipo de revestimiento de mampostería
- **MasVnrArea**: cuantitativa, área del revestimiento de mampostería en pies cuadrados
- ExterQual: cualitativa, calidad del material exterior
- ExterCond: cualitativa, condición actual del material en el exterior
- Foundation: cualitativa, tipo de cimentación
- BsmtQual: cualitativa, altura del sótano
- BsmtCond: cualitativa, condición general del sótano
- BsmtExposure: cualitativa, paredes del sótano de salida directa o a nivel de jardín
- BsmtFinType1: cualitativa, calidad del área terminada del sótano (tipo 1)
- **BsmtFinSF1**: cuantitativa, pies cuadrados terminados, tipo 1
- BsmtFinType2: cualitativa, calidad del área terminada del sótano (tipo 2, si está presente)
- **BsmtFinSF2**: cuantitativa, pies cuadrados terminados, tipo 2
- **BsmtUnfSF**: cuantitativa, pies cuadrados sin terminar del área del sótano
- **TotalBsmtSF**: cuantitativa, total de pies cuadrados del área del sótano
- Heating: cualitativa, tipo de calefacción
- HeatingQC: cualitativa, calidad y condición de la calefacción
- CentralAir: cualitativa, aire acondicionado central
- Electrical: cualitativa, sistema eléctrico
- **X1stFlrSF**: cuantitativa, pies cuadrados del primer piso
- **X2ndFlrSF**: cuantitativa, pies cuadrados del segundo piso
- **LowQualFinSF**: cuantitativa, pies cuadrados terminados de baja calidad (todas las plantas)
- **GrLivArea**: cuantitativa, pies cuadrados de área habitable sobre el nivel del suelo
- **BsmtFullBath**: cuantitativa, baños completos en el sótano
- **BsmtHalfBath**: cuantitativa, medios baños en el sótano
- **FullBath**: cuantitativa, baños completos sobre el nivel del suelo
- **HalfBath**: cuantitativa, medios baños sobre el nivel del suelo
- **BedroomAbvGr**: cuantitativa, número de dormitorios sobre el nivel del sótano
- **KitchenAbvGr**: cuantitativa, número de cocinas
- KitchenQual: cualitativa, calidad de la cocina
- **TotRmsAbvGrd**: cuantitativa, total de habitaciones sobre el nivel del suelo (no incluye baños)

- **Functional:** cualitativa, calificación de funcionalidad de la vivienda
- **Fireplaces:** cuantitativa, número de chimeneas
- **FireplaceQu:** cualitativa, calidad de la chimenea
- **GarageType:** cualitativa, ubicación del garaje
- **GarageYrBlt:** cuantitativa, año en que se construyó el garaje
- **GarageFinish:** cualitativa, acabado interior del garaje
- **GarageCars:** cuantitativa, capacidad del garaje en número de autos
- **GarageArea:** cuantitativa, tamaño del garaje en pies cuadrados
- **GarageQual:** cualitativa, calidad del garaje
- **GarageCond:** cualitativa, condición del garaje
- **PavedDrive:** cualitativa, acceso pavimentado
- **WoodDeckSF:** cuantitativa, área de la terraza de madera en pies cuadrados
- **OpenPorchSF:** cuantitativa, área del porche abierto en pies cuadrados
- **EnclosedPorch:** cuantitativa, área del porche cerrado en pies cuadrados
- **X3SsnPorch:** cuantitativa, área del porche de tres estaciones en pies cuadrados
- **ScreenPorch:** cuantitativa, área del porche con malla en pies cuadrados
- **PoolArea:** cuantitativa, área de la piscina en pies cuadrados
- **PoolQC:** cualitativa, calidad de la piscina
- **Fence:** cualitativa, calidad de la cerca
- **MiscFeature:** cualitativa, característica miscelánea no cubierta en otras categorías
- **MiscVal:** cuantitativa, valor en dólares de la característica miscelánea
- **MoSold:** cuantitativa, mes en que se vendió
- **YrSold:** cuantitativa, año en que se vendió
- **SaleType:** cualitativa, tipo de venta
- **SaleCondition:** cuantitativa, condición de la venta

Estadísticas descriptivas

##	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      :    4
3.49
##                                     3rd Qu.:
0.00
##                                     Max.      :1550
0.00
##
##      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
##

```

Variables numéricas

Estadísticas descriptivas:

```

##      SalePrice      LotFrontage      LotArea      OverallQual
##  Min.   : 34900   Min.    : 21.00   Min.    : 1300   Min.    : 1.000
## 1st Qu.:129975   1st Qu.: 59.00   1st Qu.: 7554   1st Qu.: 5.000
## Median :163000   Median : 69.00   Median : 9478   Median : 6.000
## Mean   :180921   Mean    : 70.05   Mean    :10517   Mean    : 6.099
## 3rd Qu.:214000   3rd Qu.: 80.00   3rd Qu.:11602   3rd Qu.: 7.000
## Max.   :755000   Max.    :313.00   Max.    :215245   Max.    :10.000
##
##      OverallCond      YearBuilt      YearRemodAdd      MasVnrArea
##  Min.   :1.000   Min.    :1872   Min.    :1950   Min.    : 0.0
## 1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   1st Qu.: 0.0
## Median :5.000   Median :1973   Median :1994   Median : 0.0
## Mean   :5.575   Mean    :1971   Mean    :1985   Mean    :103.7
## 3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   3rd Qu.:166.0
## Max.   :9.000   Max.    :2010   Max.    :2010   Max.    :1600.0
##
##                                     NA's      :8
##      BsmtFinSF1      BsmtFinSF2      BsmtUnfSF      TotalBsmtSF
##  Min.   : 0.0   Min.    : 0.00   Min.    : 0.0   Min.    : 0.0
## 1st Qu.: 0.0   1st Qu.: 0.00   1st Qu.:223.0   1st Qu.:795.8
## Median :383.5   Median : 0.00   Median :477.5   Median :991.5
## Mean   :443.6   Mean    :46.55   Mean    :567.2   Mean   :1057.4
## 3rd Qu.:712.2   3rd Qu.: 0.00   3rd Qu.:808.0   3rd Qu.:1298.2

```

```

## Max. :5644.0 Max. :1474.00 Max. :2336.0 Max. :6110.0
##
## X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea
## Min. : 334 Min. : 0 Min. : 0.000 Min. : 334
## 1st Qu.: 882 1st Qu.: 0 1st Qu.: 0.000 1st Qu.:1130
## Median :1087 Median : 0 Median : 0.000 Median :1464
## Mean :1163 Mean : 347 Mean : 5.845 Mean :1515
## 3rd Qu.:1391 3rd Qu.: 728 3rd Qu.: 0.000 3rd Qu.:1777
## Max. :4692 Max. :2065 Max. :572.000 Max. :5642
##
## BsmtFullBath BsmtHalfBath FullBath HalfBath
## Min. :0.0000 Min. :0.00000 Min. :0.000 Min. :0.0000
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:1.000 1st Qu.:0.0000
## Median :0.0000 Median :0.00000 Median :2.000 Median :0.0000
## Mean :0.4253 Mean :0.05753 Mean :1.565 Mean :0.3829
## 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:2.000 3rd Qu.:1.0000
## Max. :3.0000 Max. :2.00000 Max. :3.000 Max. :2.0000
##
## BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces
## Min. :0.000 Min. :0.000 Min. : 2.000 Min. :0.000
## 1st Qu.:2.000 1st Qu.:1.000 1st Qu.: 5.000 1st Qu.:0.000
## Median :3.000 Median :1.000 Median : 6.000 Median :1.000
## Mean :2.866 Mean :1.047 Mean : 6.518 Mean :0.613
## 3rd Qu.:3.000 3rd Qu.:1.000 3rd Qu.: 7.000 3rd Qu.:1.000
## Max. :8.000 Max. :3.000 Max. :14.000 Max. :3.000
##
## GarageYrBlt GarageCars GarageArea WoodDeckSF
## Min. :1900 Min. :0.000 Min. : 0.0 Min. : 0.00
## 1st Qu.:1961 1st Qu.:1.000 1st Qu.: 334.5 1st Qu.: 0.00
## Median :1980 Median :2.000 Median : 480.0 Median : 0.00
## Mean :1979 Mean :1.767 Mean : 473.0 Mean : 94.24
## 3rd Qu.:2002 3rd Qu.:2.000 3rd Qu.: 576.0 3rd Qu.:168.00
## Max. :2010 Max. :4.000 Max. :1418.0 Max. :857.00
## NA's :81
## OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00
## Median : 25.00 Median : 0.00 Median : 0.00 Median : 0.00
## Mean : 46.66 Mean : 21.95 Mean : 3.41 Mean : 15.06
## 3rd Qu.: 68.00 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 0.00
## Max. :547.00 Max. :552.00 Max. :508.00 Max. :480.00
##
## PoolArea MiscVal MoSold YrSold
## Min. : 0.000 Min. : 0.00 Min. : 1.000 Min. :2006
## 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 5.000 1st Qu.:2007
## Median : 0.000 Median : 0.00 Median : 6.000 Median :2008
## Mean : 2.759 Mean : 43.49 Mean : 6.322 Mean :2008
## 3rd Qu.: 0.000 3rd Qu.: 0.00 3rd Qu.: 8.000 3rd Qu.:2009
## Max. :738.000 Max. :15500.00 Max. :12.000 Max. :2010
##

```


Podemos notar que la gran mayoría de variables tienen datos en todas las filas. Excepto para algunos casos que a continuación se explica como se llenaran los datos faltantes:

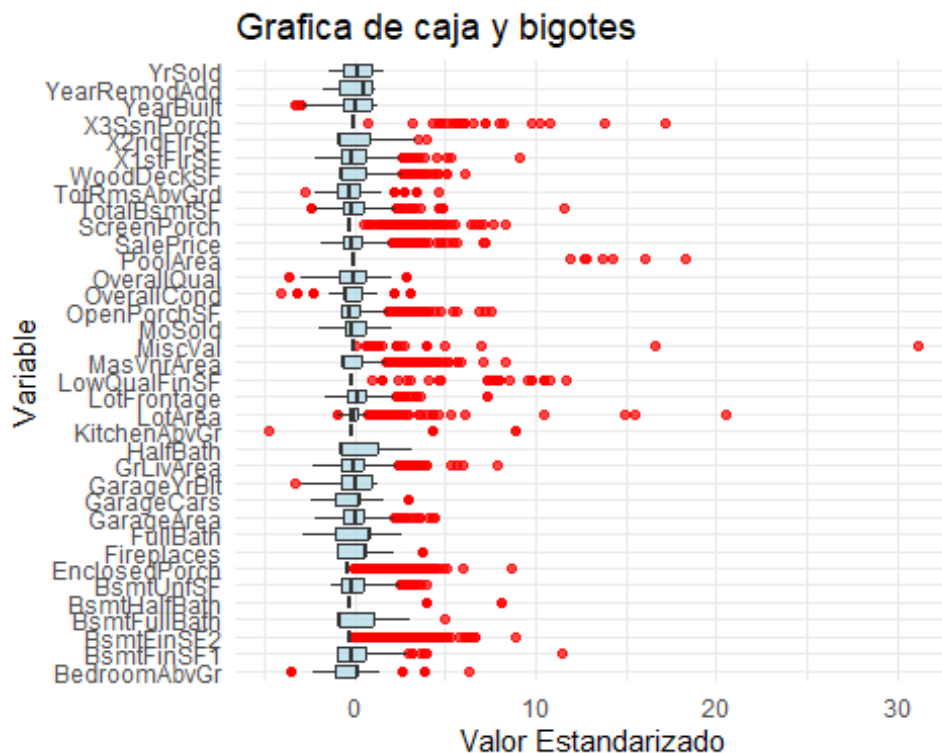
- LotFrontage: pies lineales de calle conectada a la propiedad, si hay un Na se va a considerar que la propiedad no tiene acceso directo a la calle (por ejemplo tiene derecho de paso con otra propiedad que si tiene acceso a la calle). por lo que tiene sentido remplazar los Na's por 0.
- MasVnrArea : por razon similar tiene sentido remplazar los Na's por 0
- GarageYrBlt : el año de construccion del garaje, en este caso el Na's tiene sentido porque si una casa no tiene garaje, este no tiene año de construccion. Pero para no eliminar las casas sin garaje se va a remplazar los Na's por la mediana de la columna por ser un estadístico mas robusto que la media, es decir que no tiende a sesgar los datos.

variables completadas

##	LotFrontage	MasVnrArea	GarageYrBlt
##	Min. : 0.00	Min. : 0.0	Min. :1900
##	1st Qu.: 42.00	1st Qu.: 0.0	1st Qu.:1962
##	Median : 63.00	Median : 0.0	Median :1980
##	Mean : 57.62	Mean : 103.1	Mean :1979
##	3rd Qu.: 79.00	3rd Qu.: 164.2	3rd Qu.:2001
##	Max. :313.00	Max. :1600.0	Max. :2010

Otra cosa que se puede notar al leer los resultados del summary es que algunas variables como MiscVal, teniendo una media de 43 y un tercer cuartil de 0.00 tienen probablemente muchos datos atípicos de gran valor que sesgan la media. Vamos a normalizar los datos para poder hacer diagramas de caja y bigotes de todas variables a la vez. Para analizar visualmente el summary.

Analisis de grafica de caja y bigotes



De esta esta grafica podemos notar que hay variables demasiado dispersas que podrían no ser buenas para elaborar una regresion. Otras que tienen casi todos los valores nulos. Tomando en cuenta que son variables de casas, puede ser que mansiones o casas muy grandes tengan otra escala de valores. Puede ser una buena idea crear clusters, alguna funcion que le asigne un cluster a una casa dada y crear regresiones lineales diferentes para cada cluster. Otra opcion sería no tomar en cuenta esas variables, quitar los datos atipicos y crear un modelo que sea bueno prediciendo el valor de una propiedad “usual”. Para calcular el valor de mansiones o similares puede que el modelo no vaya a ver bueno. Para indagar más en estas opciones se hara un analisis de correlacion y un clustering.

Prueba de normalidad en las variables cuantitativas

Tabla de resultados de la prueba de Shapiro-Wilk

##	Variable	P_Value
##	SalePrice	3.206142e-33
##	LotFrontage	9.099048e-30
##	LotArea	7.933654e-58
##	OverallQual	2.686457e-22
##	OverallCond	6.774229e-37
##	YearBuilt	2.770220e-26
##	YearRemodAdd	6.720280e-34
##	MasVnrArea	4.382067e-48

```
## BsmtFinSF1      BsmtFinSF1 2.813854e-35
## BsmtFinSF2      BsmtFinSF2 1.850254e-58
## BsmtUnfSF       BsmtUnfSF 1.639911e-25
## TotalBsmtSF     TotalBsmtSF 1.611332e-27
## X1stFlrSF       X1stFlrSF 4.513223e-26
## X2ndFlrSF       X2ndFlrSF 2.514882e-41
## LowQualFinSF    LowQualFinSF 9.589248e-64
## GrLivArea       GrLivArea 6.597611e-26
## BsmtFullBath    BsmtFullBath 3.760666e-47
## BsmtHalfBath    BsmtHalfBath 1.466616e-60
## FullBath        FullBath 4.231488e-44
## HalfBath        HalfBath 4.581582e-48
## BedroomAbvGr    BedroomAbvGr 4.115551e-35
## KitchenAbvGr    KitchenAbvGr 4.221203e-61
## TotRmsAbvGrd    TotRmsAbvGrd 2.004964e-23
## Fireplaces      Fireplaces 4.830980e-42
## GarageYrBlt     GarageYrBlt 7.080806e-26
## GarageCars      GarageCars 2.301685e-36
## GarageArea      GarageArea 4.016963e-15
## WoodDeckSF      WoodDeckSF 3.227985e-41
## OpenPorchSF     OpenPorchSF 1.135905e-43
## EnclosedPorch   EnclosedPorch 4.849485e-56
## X3SsnPorch      X3SsnPorch 8.307268e-64
## ScreenPorch     ScreenPorch 3.305688e-59
## PoolArea        PoolArea 7.111538e-65
## MiscVal         MiscVal 1.529907e-64
## MoSold          MoSold 3.178973e-17
## YrSold          YrSold 3.420194e-30
```

Se puede constatar que todas las variables tienen $p - valor < 0.001$, por lo que se puede concluir que ninguna de las variables sigue una distribución normal. Pero para un modelo de regresión lineal no hay un supuesto de normalidad en la distribución de las variables, esto no debería de ser un problema. Dicho usar k-medias para elaborar un clustering puede no ser la mejor opción.

Analisis de correlacion

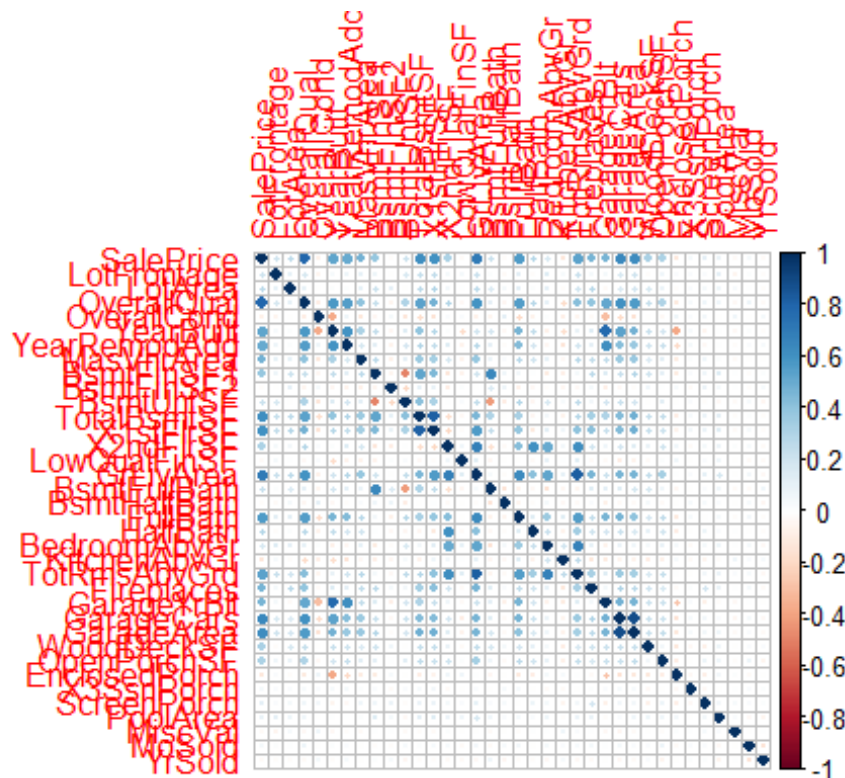
Matriz de correlacion entre variables numericas

```
rcor<-cor(datosC)
det(rcor)
```

```
## [1] -7.391762e-38
```

Notese que el determinante de la matriz de correlacion es: -7.391762e-38 por lo que hay multicolinealidad entre las variables tomadas.

```
## corplot 0.95 loaded
```



Con el mapa de calor se observa las variables relacionadas positivamente, son:

Sale price con overallQual, Year Built con GarageYrBlt, GarageCars con GarageArea, GrLivArea con TotRmsAbvGrd, 1stFlrSF con TotalBsmtSF

No se observa correlaciones negativas que sean significativas.

Las variables predictorias son las siguiente: OverallQual que tiene relacion postiva al precio de venta.. GarageArea, pues el tamano es predictor obvio, mayor area mayor precio. GarageArea, A mayor capacidad para varios carros, agrega valor la propiedad. Year Built Las casas nuevas teenden a ser mas caras. TotalBsmtSF El espacio en el sotano puede aumentar el valor

Analisis de grupos

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

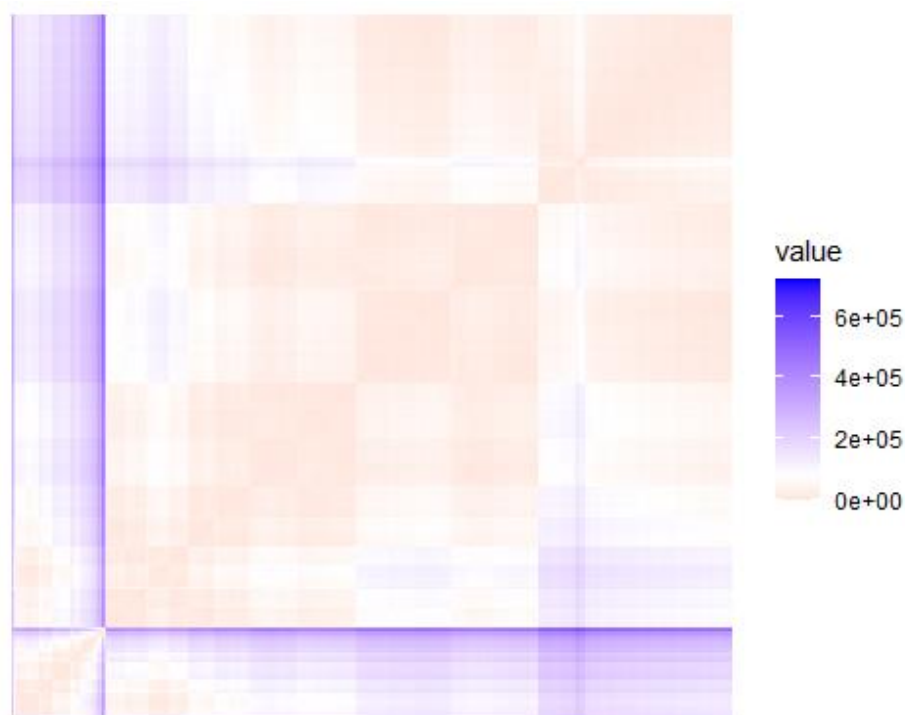
El valor estadistico de hopkins cercano a 1 indica que los datos estan altamente agrupados, por lo que vale la pena hacer agrupamiento. Para verificar, se presenta el metodo grafico.

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
datos_dist<- dist(datosC)
```

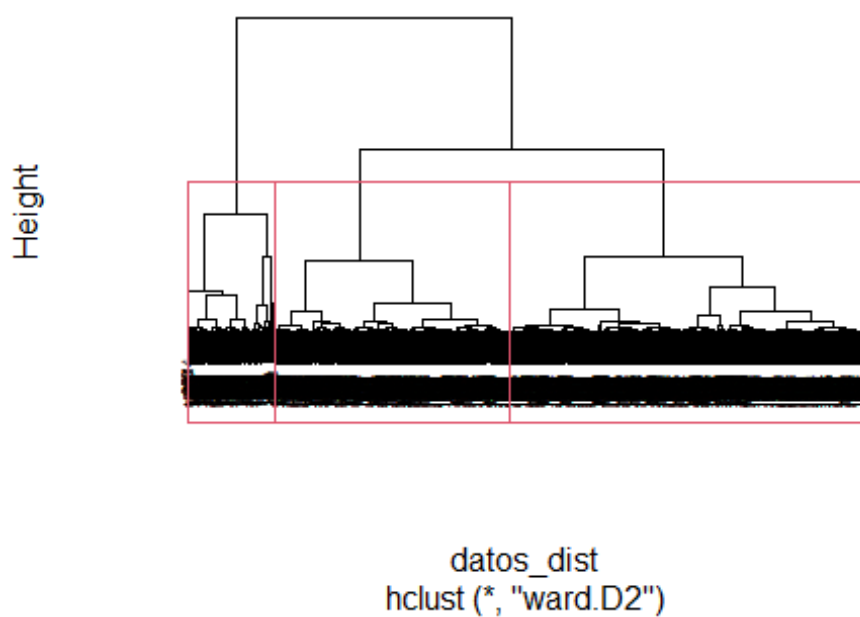
```
fviz_dist(datos_dist, show_labels=F)
```



En el mapa de calor se observan patrones de distancia que sugieren que hay tendencia a tener agrupamiento, afirmado lo que indicaba el valor estadístico de hopkins.

```
matriz<-dist(datosC)
```

Cluster Dendrogram



```

gr<-cutree(hc, k=3)
datosC$gruposHc<-gr

table(datosC$gruposHc)

##
##    1    2    3
## 509 763 188

print(table(datosC$gruposHc))

##
##    1    2    3
## 509 763 188

by(datosC, datosC[, "gruposHc"], colMeans)

## datosC[, "gruposHc"]: 1
##      SalePrice  LotFrontage      LotArea  OverallQual  OverallCond
## 2.037785e+05  5.731434e+01  1.105996e+04  6.677800e+00  5.445972e+00
##      YearBuilt  YearRemodAdd  MasVnrArea  BsmtFinSF1  BsmtFinSF2
## 1.986833e+03  1.994059e+03  1.110472e+02  4.378291e+02  4.631238e+01
##      BsmtUnfSF  TotalBsmtSF  X1stFlrSF  X2ndFlrSF  LowQualFinSF
## 6.641022e+02  1.148244e+03  1.236735e+03  4.526031e+02  5.186640e+00
##      GrLivArea  BsmtFullBath  BsmtHalfBath  FullBath  HalfBath
## 1.694525e+03  4.243615e-01  5.697446e-02  1.878193e+00  5.304519e-01
## BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd  Fireplaces  GarageYrBlt
## 2.980354e+00  1.013752e+00  6.921415e+00  7.799607e-01  1.988815e+03
##      GarageCars  GarageArea  WoodDeckSF  OpenPorchSF  EnclosedPorch
## 2.051081e+00  5.341061e+02  1.094558e+02  6.161690e+01  1.371906e+01
##      X3SsnPorch  ScreenPorch  PoolArea  MiscVal  MoSold
## 6.198428e+00  1.823969e+01  4.430255e+00  3.159136e+01  6.343811e+00
##      YrSold      gruposHc
## 2.007817e+03  1.000000e+00
## -----
## datosC[, "gruposHc"]: 2
##      SalePrice  LotFrontage      LotArea  OverallQual  OverallCond
## 1.271006e+05  5.390039e+01  8.668258e+03  5.205767e+00  5.714286e+00
##      YearBuilt  YearRemodAdd  MasVnrArea  BsmtFinSF1  BsmtFinSF2
## 1.954999e+03  1.974587e+03  5.229489e+01  3.569266e+02  4.845085e+01
##      BsmtUnfSF  TotalBsmtSF  X1stFlrSF  X2ndFlrSF  LowQualFinSF
## 4.649528e+02  8.703303e+02  1.001332e+03  2.194653e+02  6.973788e+00
##      GrLivArea  BsmtFullBath  BsmtHalfBath  FullBath  HalfBath
## 1.227771e+03  3.617300e-01  6.290957e-02  1.239843e+00  2.385321e-01
## BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd  Fireplaces  GarageYrBlt
## 2.753604e+00  1.079948e+00  5.857143e+00  3.643512e-01  1.967069e+03
##      GarageCars  GarageArea  WoodDeckSF  OpenPorchSF  EnclosedPorch
## 1.363041e+00  3.695727e+02  6.225950e+01  2.715596e+01  3.042988e+01
##      X3SsnPorch  ScreenPorch  PoolArea  MiscVal  MoSold
## 1.570118e+00  1.061599e+01  6.290957e-01  5.467104e+01  6.195282e+00
##      YrSold      gruposHc

```

```
## 2.007840e+03 2.000000e+00
## -----
## datosC[, "gruposHc"]: 3
##      SalePrice  LotFrontage  LotArea  OverallQual  OverallCond
## 3.374677e+05 7.356915e+01 1.654878e+04 8.159574e+00 5.361702e+00
##      YearBuilt  YearRemodAdd  MasVnrArea  BsmtFinSF1  BsmtFinSF2
## 1.995154e+03 2.001691e+03 2.879096e+02 8.112979e+02 3.947340e+01
##      BsmtUnfSF  TotalBsmtSF  X1stFlrSF  X2ndFlrSF  LowQualFinSF
## 7.201277e+02 1.570899e+03 1.616601e+03 5.786277e+02 3.042553e+00
##      GrLivArea  BsmtFullBath  BsmtHalfBath  FullBath  HalfBath
## 2.198271e+03 6.861702e-01 3.723404e-02 2.037234e+00 5.691489e-01
##      BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd  Fireplaces  GarageYrBlt
## 3.015957e+00 1.000000e+00 8.106383e+00 1.170213e+00 1.997654e+03
##      GarageCars  GarageArea  WoodDeckSF  OpenPorchSF  EnclosedPorch
## 2.638298e+00 7.271649e+02 1.828723e+02 8.532447e+01 9.851064e+00
##      X3SsnPorch  ScreenPorch  PoolArea  MiscVal  MoSold
## 3.324468e+00 2.449468e+01 6.877660e+00 3.031915e+01 6.776596e+00
##      YrSold  gruposHc
## 2.007713e+03 3.000000e+00
```

En el grupo 1: casas de precio medio, con calidad general buena y area habitable moderada.

En el grupo 2: Casas mas antiguas y menor precio, con menor calidad y area habitable pequeño .

En el grupo 3: casas mas nuevas y de mayor precio, con calidad mayor y area habitable grande y garage mas grande.

Veamos la silueta

```
## [1] 0.5330381
```

Aun con valor 0.5330381 sugiriendo que la agrupacion es razonablemente bueno, hay buena cohesion y separacion razonable.

```
plot(silh, cex.names=.4, col=1:3)
```

Silhouette plot of (x = gr, dist = matriz)

n = 1460

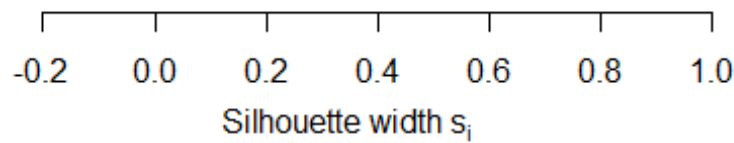
3 clusters C_j

$j: n_j | \text{ave}_{i \in C_j} s_i$

1: 509 | 0.53

2: 763 | 0.59

3: 188 | 0.32



Average silhouette width : 0.53

En la agrupacion 3 es el menos bien definido requiriendo un analisis mas detallado. La tabla de clustering resultante muestra las observaciones de cada agrupamiento

```
library(mclust)

## Package 'mclust' version 6.1.1
## Type 'citation("mclust")' for citing this R package in publications.

mc<-Mclust(datosC, 3)
summary(mc)

## -----
## Gaussian finite mixture model fitted by EM algorithm
## -----
##
## Mclust VEI (diagonal, equal shape) model with 3 components:
##
##   log-likelihood    n df         BIC          ICL
##   -234472.6 1460 152 -470052.7 -470084.6
##
## Clustering table:
##   1  2  3
## 644 211 605
```


Modelos de regresion lineal

Variable Respuesta:

La variable que se quiere predecir es el precio de las casasa, es decir (Sale Price).

Conjunto de entrenamiento y prueba:

Se usara para un 80% de los datos para entrenamiento y 20% para la prueba. Manteniendo la proporcion de observaciones de cada cluster.

```
library(caret)

## Cargando paquete requerido: lattice

#VARIABLE objetivo
y<- datosC$SalePrice
set.seed(123)
trainI<- createDataPartition(y, p=0.8, list=FALSE)
train<-datosC[trainI, ]
test<-datosC[-trainI, ]

cat("Conjunto de entrenamiento (cantidad de muestras:", nrow(train), ") \n ")

## Conjunto de entrenamiento (cantidad de muestras: 1169 )
##

head(train)

##   SalePrice LotFrontage LotArea OverallQual OverallCond YearBuilt Year
RemodAdd
## 1    208500         65    8450          7           5      2003
2003
## 2    181500         80    9600          6           8      1976
1976
## 3    223500         68   11250          7           5      2001
2002
## 4    140000         60    9550          7           5      1915
1970
## 5    250000         84   14260          8           5      2000
2000
## 6    143000         85   14115          5           5      1993
1995
##   MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF X1stFlrSF X2n
dFlrSF
## 1         196        706          0       150         856         856
854
## 2           0        978          0       284        1262        1262
0
```

## 3	162	486	0	434	920	920
866						
## 4	0	216	0	540	756	961
756						
## 5	350	655	0	490	1145	1145
1053						
## 6	0	732	0	64	796	796
566						
##	LowQualFinSF	GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath
## 1	0	1710	1	0	2	1
## 2	0	1262	0	1	2	0
## 3	0	1786	1	0	2	1
## 4	0	1717	1	0	1	0
## 5	0	2198	1	0	2	1
## 6	0	1362	1	0	1	1
##	BedroomAbvGr	KitchenAbvGr	TotRmsAbvGrd	Fireplaces	GarageYrBlt	Garage
Cars						
## 1	3	1	8	0	2003	
2						
## 2	3	1	6	1	1976	
2						
## 3	3	1	6	1	2001	
2						
## 4	3	1	7	1	1998	
3						
## 5	4	1	9	1	2000	
3						
## 6	1	1	5	0	1993	
2						
##	GarageArea	WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch	ScreenPor
ch						
## 1	548	0	61	0	0	
0						
## 2	460	298	0	0	0	
0						
## 3	608	0	42	0	0	
0						
## 4	642	0	35	272	0	
0						
## 5	836	192	84	0	0	
0						
## 6	480	40	30	0	320	
0						
##	PoolArea	MiscVal	MoSold	YrSold	gruposHc	
## 1	0	0	2	2008	1	
## 2	0	0	5	2007	1	
## 3	0	0	9	2008	1	
## 4	0	0	2	2006	2	
## 5	0	0	12	2008	1	
## 6	0	700	10	2009	2	

```

cat("Conjunto de prueba (cantidad de muestras:", nrow(test), ") \n ")

## Conjunto de prueba (cantidad de muestras: 291 )
##

head(test)

##      SalePrice LotFrontage LotArea OverallQual OverallCond YearBuilt Year
rRemodAdd
## 8      200000          0    10382          7          6      1973
1973
## 11     129500         70    11200          5          5      1965
1965
## 13     144000          0    12968          5          6      1962
1962
## 30      68500         60     6324          4          6      1927
1950
## 37     145000        112    10859          5          5      1994
1995
## 39     109000         68     7922          5          7      1953
2007
##      MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF X1stFlrSF X2
ndFlrSF
## 8          240         859          32         216         1107         1107
983
## 11           0         906           0         134         1040         1040
0
## 13           0         737           0         175          912          912
0
## 30           0           0           0         520          520          520
0
## 37           0           0           0        1097         1097         1097
0
## 39           0         731           0         326         1057         1057
0
##      LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
## 8           0         2090           1           0           2           1
## 11          0         1040           1           0           1           0
## 13          0          912           1           0           1           0
## 30          0          520           0           0           1           0
## 37          0         1097           0           0           1           1
## 39          0         1057           1           0           1           0
##      BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt Garag
eCars
## 8           3           1           7           2           1973
2
## 11          3           1           5           0           1965
1
## 13          2           1           4           0           1962
1
## 30          1           1           4           0           1920

```

```

1
## 37          3          1          6          0          1995
2
## 39          3          1          5          0          1953
1
##      GarageArea WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPo
rch
## 8          484          235          204          228          0
0
## 11          384           0           0           0           0
0
## 13          352          140           0           0           0
176
## 30          240           49           0           87           0
0
## 37          672          392           64           0           0
0
## 39          246           0           52           0           0
0
##      PoolArea MiscVal MoSold YrSold gruposHc
## 8           0       350       11   2009        1
## 11          0         0         2   2008        2
## 13          0         0         9   2008        2
## 30          0         0         5   2008        2
## 37          0         0         6   2009        2
## 39          0         0         1   2010        2

```

Modelo univariado de regresión

Primero elijamos la variable independiente para el modelos univariado, lo idóneo es elegir la variable con mayor correlación. A continuación se muestras las 5 variables con mayor correlación a SalePrice:

```

##      SalePrice OverallQual  GrLivArea  GarageCars  GarageArea
##      1.0000000    0.7909816    0.7086245    0.6404092    0.6234314

```

La variable que se utilizara para este primer modelo es OverallQual, pues es la que mayor correlacion con SalePrice tiene (sin ser SalePrice).

```

modelo0 <- lm(SalePrice~., data = train[,c("SalePrice", "OverallQual")])
summary(modelo0)

##
## Call:
## lm(formula = SalePrice ~ ., data = train[, c("SalePrice", "OverallQual
")])
##
## Residuals:
##      Min        1Q    Median        3Q        Max

```

```
## -193612 -28187 -1398 21165 273784
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -91015.7     6114.5  -14.88  <2e-16 ***
## OverallQual  44462.8      979.6   45.39  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45840 on 1167 degrees of freedom
## Multiple R-squared:  0.6384, Adjusted R-squared:  0.6381
## F-statistic: 2060 on 1 and 1167 DF, p-value: < 2.2e-16
```

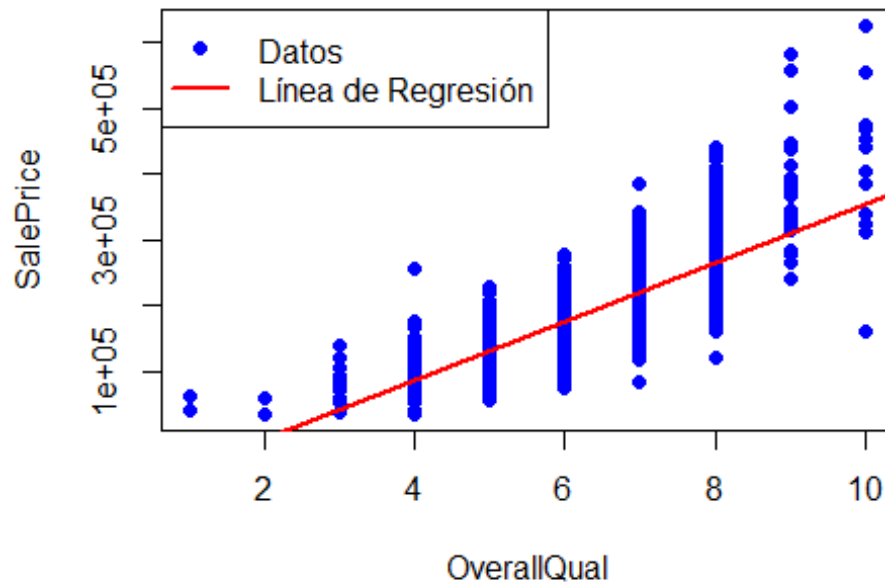
La ecuación de la regresión es: $SalePrice = -91015.7 + 44462.8OverallQual$, y el modelo explica en 64% de la varianza.

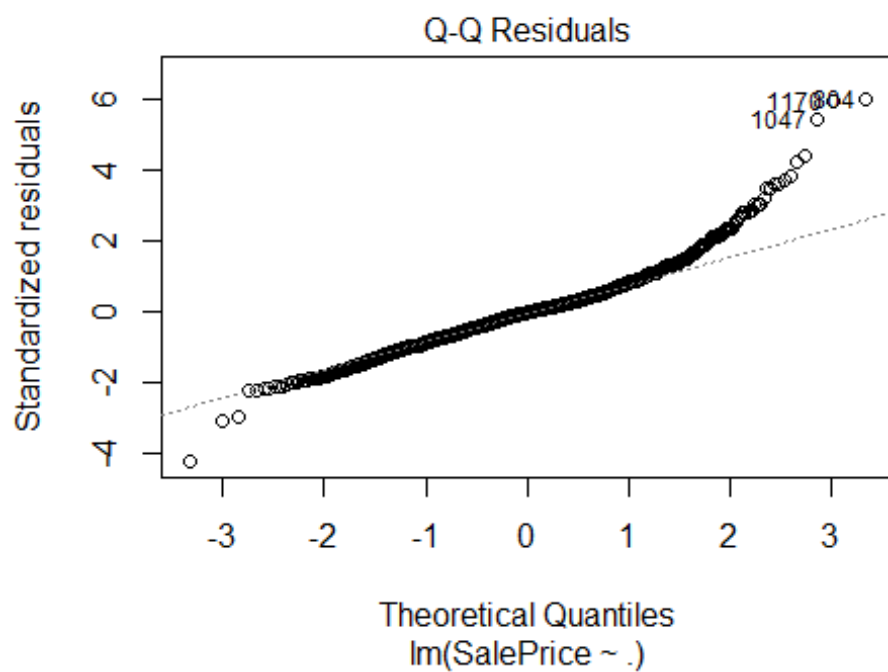
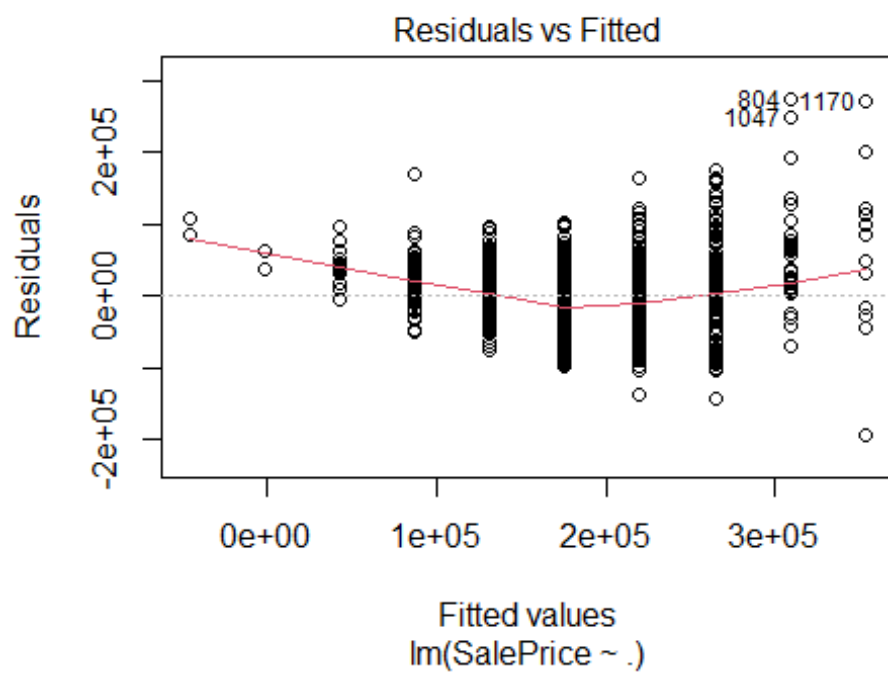
```
plot(train$OverallQual, train$SalePrice,
     main = "Regresión lineal",
     xlab = "OverallQual",
     ylab = "SalePrice",
     pch = 16, col = "blue")

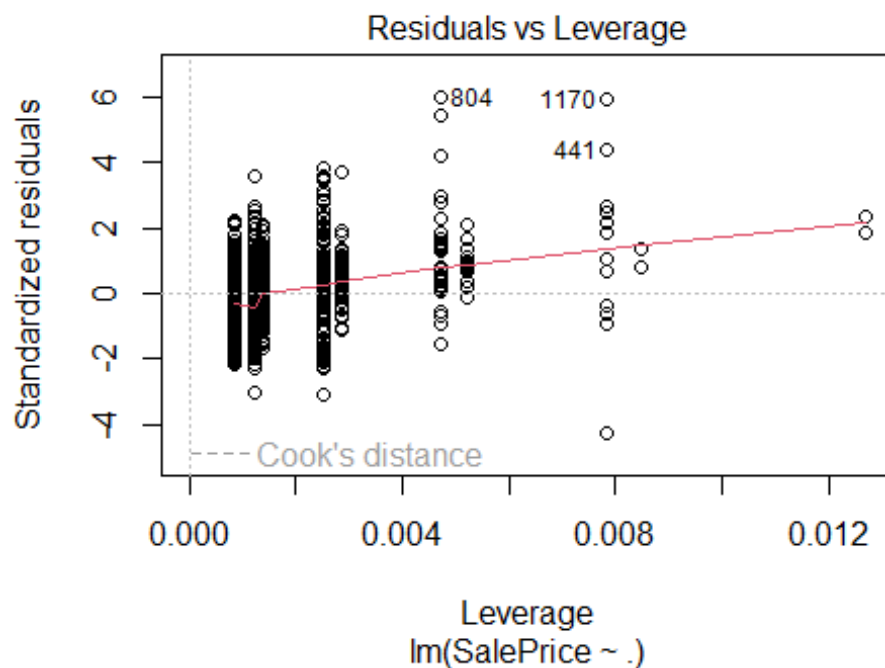
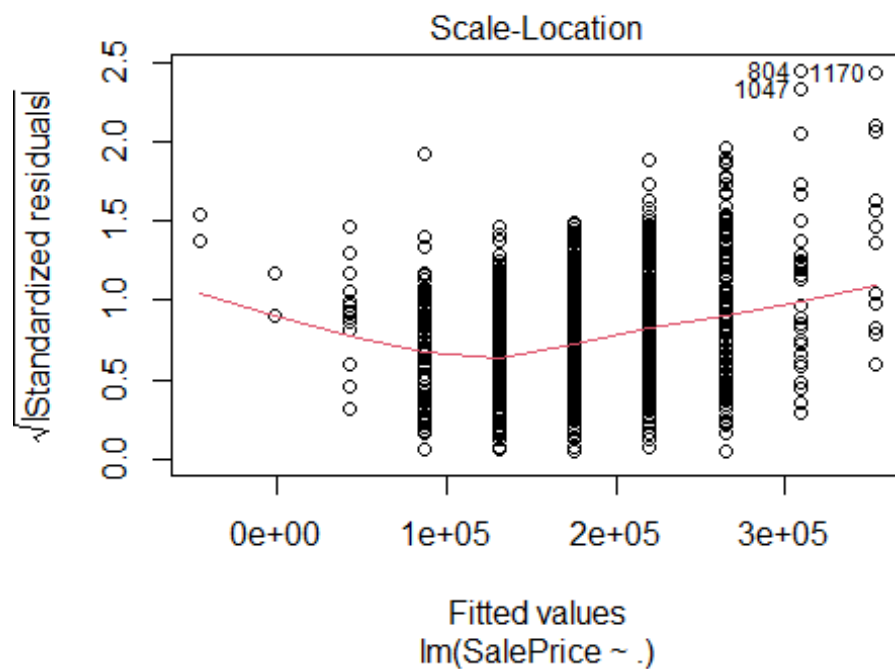
# Agrega la línea de regresión ajustada
abline(modelo0, col = "red", lwd = 2)

# Agrega una Leyenda
legend("topleft", legend = c("Datos", "Línea de Regresión"),
      col = c("blue", "red"), pch = c(16, NA), lwd = c(NA, 2))
```

Regresión lineal







Se puede ver que la varianza no es constante, En el gráfico q-q se observa que los residuos no parecen estar normalmente distribuidos. Este modelo lineal no parece ser muy bueno.


```
lillie.test(modelo0$residuals)
```

```
##  
##  Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data:  modelo0$residuals  
## D = 0.076459, p-value < 2.2e-16
```

Los residuos tampoco se distribuyen de manera normal. En resumen este modelo lineal de una variable no es bueno.

Modelo con todas las variables

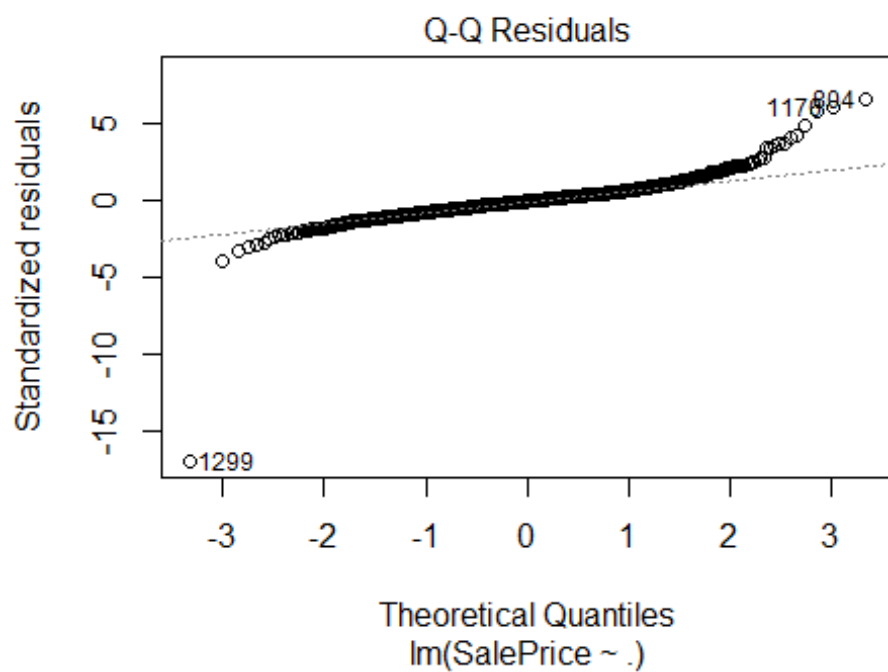
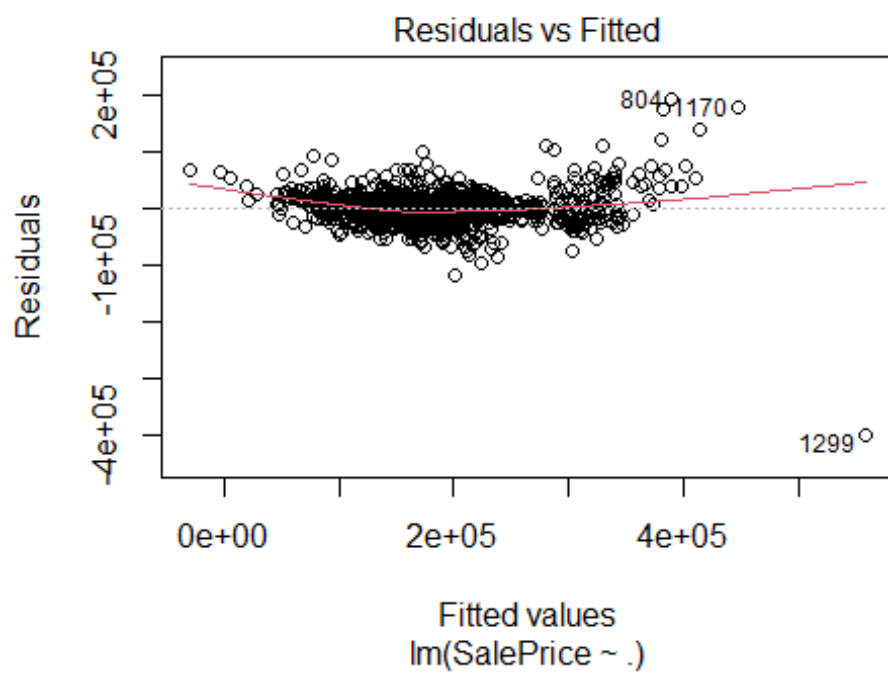
```
modelo1 <- lm(SalePrice~. ,data = train)  
summary(modelo1)
```

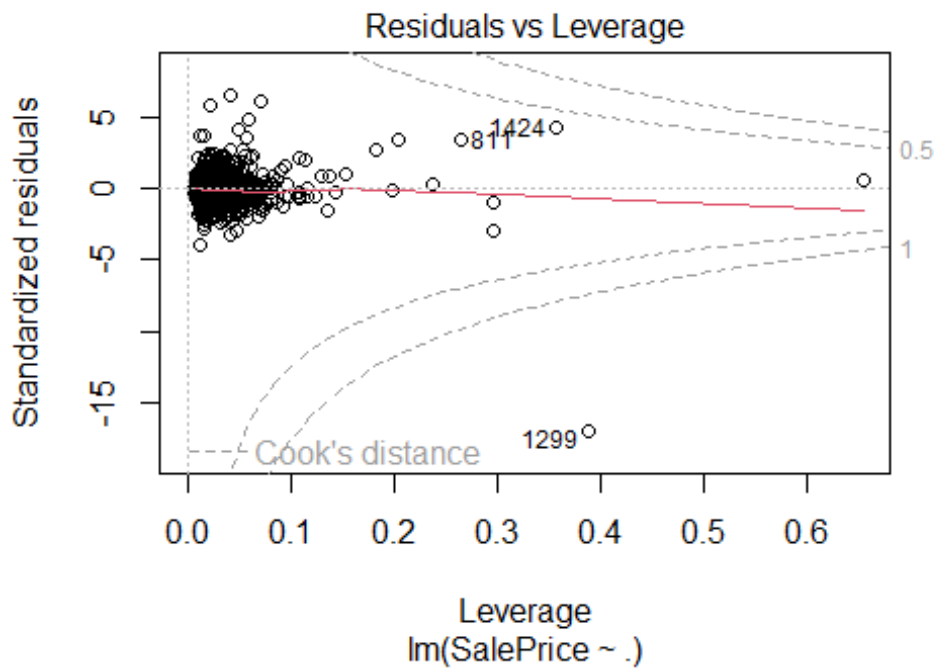
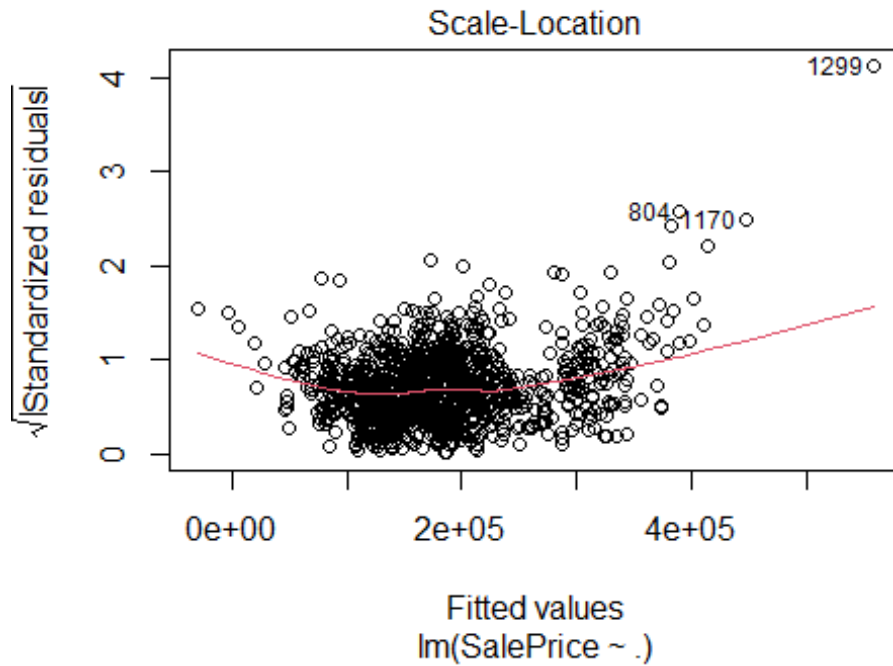
```
##  
## Call:  
## lm(formula = SalePrice ~ ., data = train)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -397446  -15279    -648   12513  193096   
##  
## Coefficients: (2 not defined because of singularities)  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  6.098e+05  1.366e+06   0.447  0.655312      
## LotFrontage  1.446e+01  2.686e+01   0.538  0.590497      
## LotArea      3.870e-01  1.092e-01   3.545  0.000409 ***  
## OverallQual  1.399e+04  1.145e+03  12.222 < 2e-16 ***  
## OverallCond  5.893e+03  9.984e+02   5.903  4.71e-09 ***  
## YearBuilt    3.905e+02  6.567e+01   5.946  3.65e-09 ***  
## YearRemodAdd 1.532e+02  6.611e+01   2.317  0.020670 *    
## MasVnrArea   1.608e+01  5.909e+00   2.720  0.006619 **    
## BsmtFinSF1   1.197e+01  4.427e+00   2.704  0.006954 **    
## BsmtFinSF2   1.033e+01  6.606e+00   1.564  0.118043      
## BsmtUnfSF    1.011e+01  4.023e+00   2.513  0.012114 *     
## TotalBsmtSF      NA         NA         NA         NA      
## X1stFlrSF     4.587e+01  5.660e+00   8.106  1.35e-15 ***  
## X2ndFlrSF     3.355e+01  4.833e+00   6.942  6.48e-12 ***  
## LowQualFinSF  2.895e+01  1.868e+01   1.550  0.121473      
## GrLivArea      NA         NA         NA         NA      
## BsmtFullBath  9.910e+03  2.511e+03   3.947  8.39e-05 ***  
## BsmtHalfBath -2.850e+02  4.172e+03  -0.068  0.945552      
## FullBath      1.026e+04  2.807e+03   3.654  0.000270 ***  
## HalfBath      2.455e+03  2.648e+03   0.927  0.354081      
## BedroomAbvGr -8.076e+03  1.628e+03  -4.962  8.03e-07 ***  
## KitchenAbvGr -2.938e+04  4.694e+03  -6.259  5.48e-10 ***  
## TotRmsAbvGrd  6.090e+03  1.221e+03   4.989  7.01e-07 ***  
## Fireplaces    4.663e+03  1.755e+03   2.656  0.008016 **
```

```
## GarageYrBlt      7.785e+01  6.686e+01   1.164  0.244545
## GarageCars       1.153e+04  2.711e+03   4.253  2.28e-05 ***
## GarageArea      -1.796e+00  9.375e+00  -0.192  0.848138
## WoodDeckSF       2.008e+01  7.844e+00   2.560  0.010604 *
## OpenPorchSF      2.686e+01  1.488e+01   1.806  0.071251 .
## EnclosedPorch    1.846e+01  1.680e+01   1.099  0.272200
## X3SsnPorch       5.161e+01  3.241e+01   1.593  0.111538
## ScreenPorch      5.300e+01  1.727e+01   3.069  0.002197 **
## PoolArea        -1.080e+02  2.319e+01  -4.659  3.56e-06 ***
## MiscVal          -2.267e+00  2.982e+00  -0.760  0.447352
## MoSold           7.637e+01  3.319e+02   0.230  0.818077
## YrSold           -9.608e+02  6.798e+02  -1.413  0.157842
## gruposHc         2.045e+04  1.474e+03  13.876  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29930 on 1134 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8457
## F-statistic: 189.2 on 34 and 1134 DF,  p-value: < 2.2e-16
```

Las variables significativas mostrando relacion con el precio de ventas, entran:
OverallQual: mas calidad, mayor precio de venta LotArea: mas area, mayor precio de
venta YearBuild: Entre mas nueva la vivienda mayor precio de venta BsmtFinSF1:
mayor pies cuadrados terminados de sotano mayor precio de venta X1stFlrSF: mayor
pies cuadrados terminados en los pisos mayor precio de venta BedroomAbvrGr: Mas
cantidad de cuartos mayor precio de venta KitchenAbvGRr: menor numero de cocinas
mayor precio de venta TotRms: Mas abitaciones mayor precio de venta GarageCars:
Mas espacio de garage mayor precio de venta WoodDeckSF: Mayor cantidad de pies
cuadrados de la cubierta de madera mayor precio de venta ScreePorch: Mayor pies
cuadrado de porche con mosquitero mayor precio de venta PoolArea: mayor area de
picina mayor precio de venta

```
plot(modelo1)
```





La mayoría de puntos están distribuidos aleatoriamente al rededor de cero, por lo que hay cierta heterocedasticidad, por dispersión de los residuales, sugiriendo que la dispersión de residuales aumenta ligeramente a medida que aumentan los valores ajustados. No respetando los supuestos de regresión lineal

```
library(nortest)
library(dplyr)
lillie.test(modelo1$residuals)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  modelo1$residuals
## D = 0.098257, p-value < 2.2e-16
```

El p-valor es menor a 0.5 por lo que se rechaza hipótesis nula, no hay distribución normal. Se normaliza

```
train_normal <- as.data.frame(scale(train))
test_normal <- as.data.frame(scale(test))

modelo1normalizado <- lm(SalePrice~. ,data = train_normal)
summary(modelo1normalizado)

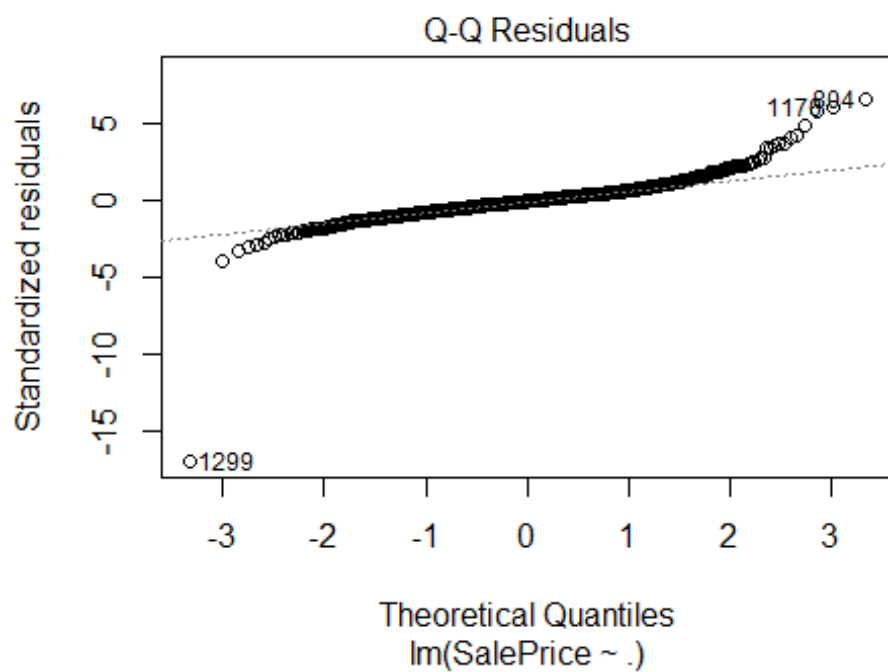
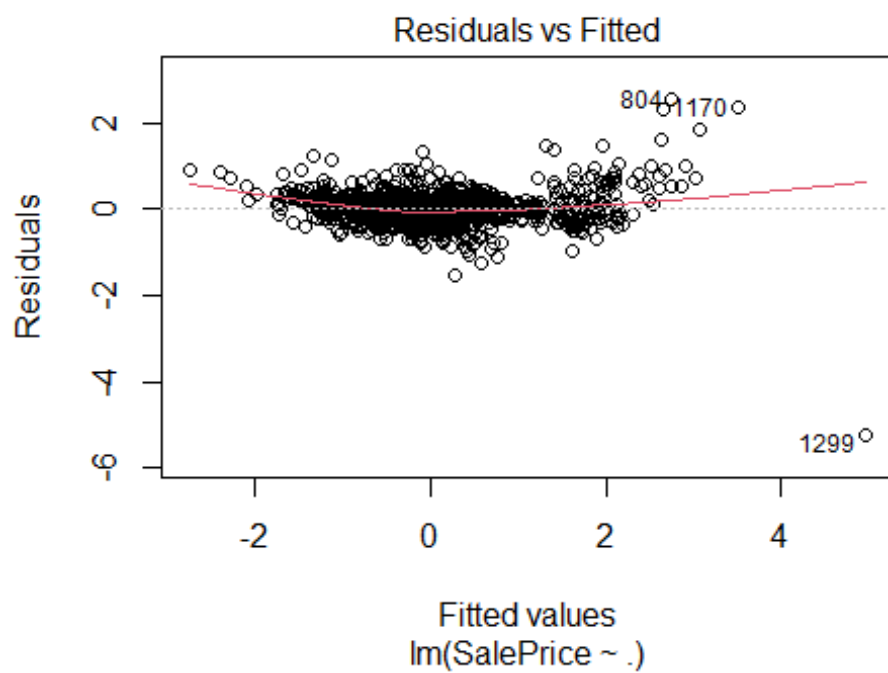
##
## Call:
## lm(formula = SalePrice ~ ., data = train_normal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.2163 -0.2005 -0.0085  0.1642  2.5343
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.683e-15  1.149e-02   0.000  1.000000
## LotFrontage   6.654e-03  1.236e-02   0.538  0.590497
## LotArea       4.584e-02  1.293e-02   3.545  0.000409 ***
## OverallQual   2.514e-01  2.057e-02  12.222 < 2e-16 ***
## OverallCond   8.561e-02  1.450e-02   5.903  4.71e-09 ***
## YearBuilt     1.550e-01  2.606e-02   5.946  3.65e-09 ***
## YearRemodAdd  4.157e-02  1.794e-02   2.317  0.020670 *
## MasVnrArea    3.737e-02  1.374e-02   2.720  0.006619 **
## BsmtFinSF1    7.158e-02  2.647e-02   2.704  0.006954 **
## BsmtFinSF2    2.282e-02  1.459e-02   1.564  0.118043
## BsmtUnfSF     5.844e-02  2.325e-02   2.513  0.012114 *
## TotalBsmtSF   NA         NA         NA         NA
## X1stFlrSF     2.297e-01  2.834e-02   8.106  1.35e-15 ***
## X2ndFlrSF     1.889e-01  2.721e-02   6.942  6.48e-12 ***
## LowQualFinSF  1.901e-02  1.226e-02   1.550  0.121473
## GrLivArea     NA         NA         NA         NA
## BsmtFullBath   6.741e-02  1.708e-02   3.947  8.39e-05 ***
## BsmtHalfBath  -8.463e-04  1.239e-02  -0.068  0.945552
## FullBath       7.314e-02  2.001e-02   3.654  0.000270 ***
## HalfBath       1.609e-02  1.736e-02   0.927  0.354081
## BedroomAbvGr  -8.788e-02  1.771e-02  -4.962  8.03e-07 ***
## KitchenAbvGr  -8.488e-02  1.356e-02  -6.259  5.48e-10 ***
```

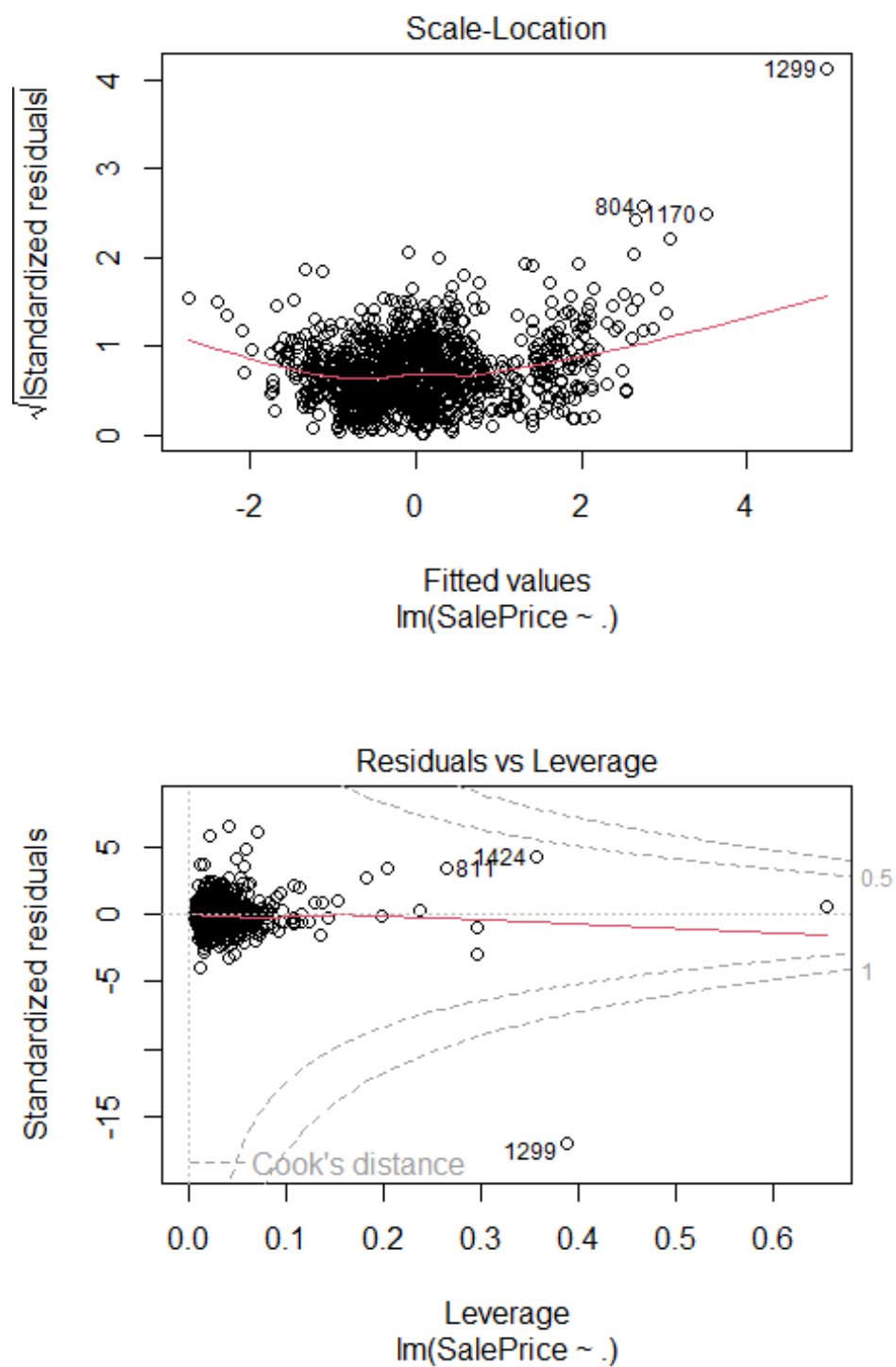
```

## TotRmsAbvGrd    1.280e-01  2.565e-02   4.989 7.01e-07 ***
## Fireplaces      3.915e-02  1.474e-02   2.656 0.008016 **
## GarageYrBlt     2.441e-02  2.097e-02   1.164 0.244545
## GarageCars      1.137e-01  2.673e-02   4.253 2.28e-05 ***
## GarageArea     -5.081e-03  2.653e-02  -0.192 0.848138
## WoodDeckSF      3.306e-02  1.292e-02   2.560 0.010604 *
## OpenPorchSF     2.296e-02  1.271e-02   1.806 0.071251 .
## EnclosedPorch   1.444e-02  1.315e-02   1.099 0.272200
## X3SsnPorch      1.865e-02  1.171e-02   1.593 0.111538
## ScreenPorch     3.787e-02  1.234e-02   3.069 0.002197 **
## PoolArea       -5.548e-02  1.191e-02  -4.659 3.56e-06 ***
## MiscVal        -9.074e-03  1.194e-02  -0.760 0.447352
## MoSold          2.733e-03  1.188e-02   0.230 0.818077
## YrSold         -1.673e-02  1.184e-02  -1.413 0.157842
## gruposHc        1.754e-01  1.264e-02  13.876 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3929 on 1134 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8457
## F-statistic: 189.2 on 34 and 1134 DF,  p-value: < 2.2e-16

plot(modelo1normalizado)

```





Notese que normalizando los datos no mejoro significativamente, por lo que se selecciona otros predictores.

Regresion con variables significativas

Vamos a hacer un ultimo modelo seleccionando solo variables significativas y que no tengan correlacion entre ellas,

```
library(glmnet)

## Cargando paquete requerido: Matrix

##
## Adjuntando el paquete: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-8

# Prepara las variables independientes (X) y la dependiente (y)
X <- model.matrix(SalePrice ~ . , data = train)[, -1] # Matriz sin el in
tercepto
y <- train$SalePrice

set.seed(123)
modelo_lasso <- cv.glmnet(X, y, alpha = 1) # alpha = 1 para Lasso

# Muestra el mejor lambda según validación cruzada
mejor_lambda <- modelo_lasso$lambda.min
cat("Mejor lambda:", mejor_lambda, "\n")

## Mejor lambda: 1341.848

coeficientes <- coef(modelo_lasso, s = "lambda.min")
variables_seleccionadas <- rownames(coeficientes)[coeficientes[, 1] != 0]
variables_seleccionadas <- variables_seleccionadas[-1] # Excluye el inte
rcepto

cat("Variables seleccionadas:", variables_seleccionadas, "\n")

## Variables seleccionadas: LotArea OverallQual OverallCond YearBuilt Yea
rRemodAdd MasVnrArea BsmtFinSF1 TotalBsmtSF X1stFlrSF GrLivArea BsmtFullB
ath FullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBl
t GarageCars GarageArea WoodDeckSF OpenPorchSF X3SsnPorch ScreenPorch Poo
lArea gruposHc

# Filtra las variables seleccionadas en el data frame
formula <- as.formula(paste("SalePrice ~", paste(variables_seleccionadas,
collapse = "+")))
modelo2 <- lm(formula, data = train)
modelo2 <- step(modelo2, direction = "backward")
```

```

## Start: AIC=24118.89
## SalePrice ~ LotArea + OverallQual + OverallCond + YearBuilt +
##      YearRemodAdd + MasVnrArea + BsmtFinSF1 + TotalBsmtSF + X1stFlrSF +
##      GrLivArea + BsmtFullBath + FullBath + BedroomAbvGr + KitchenAbvGr
+
##      TotRmsAbvGrd + Fireplaces + GarageYrBlt + GarageCars + GarageArea
+
##      WoodDeckSF + OpenPorchSF + X3SsnPorch + ScreenPorch + PoolArea +
##      gruposHc
##
##              Df  Sum of Sq      RSS   AIC
## - GarageArea    1 4.4945e+07 1.0207e+12 24117
## - BsmtFinSF1    1 2.9476e+08 1.0210e+12 24117
## - GarageYrBlt   1 9.2115e+08 1.0216e+12 24118
## <none>                                1.0207e+12 24119
## - X3SsnPorch    1 2.1892e+09 1.0229e+12 24119
## - OpenPorchSF   1 3.3584e+09 1.0240e+12 24121
## - X1stFlrSF     1 4.7229e+09 1.0254e+12 24122
## - YearRemodAdd  1 4.9875e+09 1.0257e+12 24123
## - WoodDeckSF    1 5.2956e+09 1.0260e+12 24123
## - TotalBsmtSF   1 6.3247e+09 1.0270e+12 24124
## - Fireplaces    1 6.4614e+09 1.0271e+12 24124
## - MasVnrArea    1 6.7634e+09 1.0274e+12 24125
## - ScreenPorch   1 8.0084e+09 1.0287e+12 24126
## - LotArea       1 1.0988e+10 1.0317e+12 24129
## - FullBath      1 1.1621e+10 1.0323e+12 24130
## - BsmtFullBath  1 1.5954e+10 1.0366e+12 24135
## - GarageCars    1 1.7684e+10 1.0384e+12 24137
## - PoolArea      1 1.9413e+10 1.0401e+12 24139
## - BedroomAbvGr  1 2.2043e+10 1.0427e+12 24142
## - TotRmsAbvGrd  1 2.2384e+10 1.0431e+12 24142
## - OverallCond   1 2.9671e+10 1.0503e+12 24150
## - KitchenAbvGr  1 3.7924e+10 1.0586e+12 24160
## - YearBuilt     1 3.9746e+10 1.0604e+12 24162
## - GrLivArea     1 6.7625e+10 1.0883e+12 24192
## - OverallQual   1 1.4008e+11 1.1608e+12 24267
## - gruposHc      1 1.7669e+11 1.1974e+12 24304
##
## Step: AIC=24116.94
## SalePrice ~ LotArea + OverallQual + OverallCond + YearBuilt +
##      YearRemodAdd + MasVnrArea + BsmtFinSF1 + TotalBsmtSF + X1stFlrSF +
##      GrLivArea + BsmtFullBath + FullBath + BedroomAbvGr + KitchenAbvGr
+
##      TotRmsAbvGrd + Fireplaces + GarageYrBlt + GarageCars + WoodDeckSF
+
##      OpenPorchSF + X3SsnPorch + ScreenPorch + PoolArea + gruposHc
##
##              Df  Sum of Sq      RSS   AIC
## - BsmtFinSF1    1 2.8130e+08 1.0210e+12 24115
## - GarageYrBlt   1 8.7637e+08 1.0216e+12 24116

```

```

## <none> 1.0207e+12 24117
## - X3SsnPorch 1 2.1812e+09 1.0229e+12 24117
## - OpenPorchSF 1 3.3198e+09 1.0240e+12 24119
## - X1stFlrSF 1 4.6781e+09 1.0254e+12 24120
## - YearRemodAdd 1 5.1389e+09 1.0259e+12 24121
## - WoodDeckSF 1 5.3146e+09 1.0260e+12 24121
## - TotalBsmtSF 1 6.2944e+09 1.0270e+12 24122
## - Fireplaces 1 6.6209e+09 1.0273e+12 24123
## - MasVnrArea 1 6.7332e+09 1.0275e+12 24123
## - ScreenPorch 1 7.9864e+09 1.0287e+12 24124
## - LotArea 1 1.0944e+10 1.0317e+12 24127
## - FullBath 1 1.1847e+10 1.0326e+12 24128
## - BsmtFullBath 1 1.5932e+10 1.0367e+12 24133
## - PoolArea 1 1.9475e+10 1.0402e+12 24137
## - BedroomAbvGr 1 2.2002e+10 1.0427e+12 24140
## - TotRmsAbvGrd 1 2.2451e+10 1.0432e+12 24140
## - OverallCond 1 2.9709e+10 1.0504e+12 24149
## - KitchenAbvGr 1 3.7880e+10 1.0586e+12 24158
## - YearBuilt 1 3.9964e+10 1.0607e+12 24160
## - GarageCars 1 4.5165e+10 1.0659e+12 24166
## - GrLivArea 1 6.7902e+10 1.0886e+12 24190
## - OverallQual 1 1.4027e+11 1.1610e+12 24266
## - gruposHc 1 1.7683e+11 1.1976e+12 24302
##
## Step: AIC=24115.26
## SalePrice ~ LotArea + OverallQual + OverallCond + YearBuilt +
## YearRemodAdd + MasVnrArea + TotalBsmtSF + X1stFlrSF + GrLivArea +
## BsmtFullBath + FullBath + BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd +
## Fireplaces + GarageYrBlt + GarageCars + WoodDeckSF + OpenPorchSF +
## X3SsnPorch + ScreenPorch + PoolArea + gruposHc
##
## Df Sum of Sq RSS AIC
## - GarageYrBlt 1 8.5510e+08 1.0219e+12 24114
## <none> 1.0210e+12 24115
## - X3SsnPorch 1 2.2525e+09 1.0233e+12 24116
## - OpenPorchSF 1 3.2867e+09 1.0243e+12 24117
## - X1stFlrSF 1 4.6713e+09 1.0257e+12 24119
## - YearRemodAdd 1 5.0229e+09 1.0260e+12 24119
## - WoodDeckSF 1 5.2746e+09 1.0263e+12 24119
## - Fireplaces 1 6.8593e+09 1.0279e+12 24121
## - MasVnrArea 1 7.1634e+09 1.0282e+12 24121
## - TotalBsmtSF 1 7.3882e+09 1.0284e+12 24122
## - ScreenPorch 1 7.9499e+09 1.0290e+12 24122
## - LotArea 1 1.1091e+10 1.0321e+12 24126
## - FullBath 1 1.1809e+10 1.0328e+12 24127
## - PoolArea 1 1.9272e+10 1.0403e+12 24135
## - TotRmsAbvGrd 1 2.2222e+10 1.0432e+12 24138
## - BedroomAbvGr 1 2.2428e+10 1.0434e+12 24139
## - BsmtFullBath 1 2.7385e+10 1.0484e+12 24144

```

```

## - OverallCond      1 3.0761e+10 1.0518e+12 24148
## - KitchenAbvGr     1 3.8037e+10 1.0590e+12 24156
## - YearBuilt         1 4.1169e+10 1.0622e+12 24160
## - GarageCars        1 4.5020e+10 1.0660e+12 24164
## - GrLivArea         1 6.9766e+10 1.0908e+12 24191
## - OverallQual       1 1.4066e+11 1.1617e+12 24264
## - gruposHc          1 1.7993e+11 1.2009e+12 24303
##
## Step:  AIC=24114.24
## SalePrice ~ LotArea + OverallQual + OverallCond + YearBuilt +
##      YearRemodAdd + MasVnrArea + TotalBsmtSF + X1stFlrSF + GrLivArea +
##      BsmtFullBath + FullBath + BedroomAbvGr + KitchenAbvGr + TotRmsAbvGr
rd +
##      Fireplaces + GarageCars + WoodDeckSF + OpenPorchSF + X3SsnPorch +
##      ScreenPorch + PoolArea + gruposHc
##
##              Df  Sum of Sq      RSS   AIC
## <none>                1.0219e+12 24114
## - X3SsnPorch          1 2.2562e+09 1.0241e+12 24115
## - OpenPorchSF         1 3.5849e+09 1.0254e+12 24116
## - X1stFlrSF           1 4.4787e+09 1.0263e+12 24117
## - WoodDeckSF          1 5.6942e+09 1.0276e+12 24119
## - Fireplaces          1 6.2978e+09 1.0282e+12 24119
## - YearRemodAdd        1 6.6038e+09 1.0285e+12 24120
## - MasVnrArea          1 7.0995e+09 1.0290e+12 24120
## - TotalBsmtSF         1 7.5564e+09 1.0294e+12 24121
## - ScreenPorch         1 7.9246e+09 1.0298e+12 24121
## - LotArea             1 1.0806e+10 1.0327e+12 24125
## - FullBath            1 1.2745e+10 1.0346e+12 24127
## - PoolArea            1 1.9248e+10 1.0411e+12 24134
## - TotRmsAbvGrd       1 2.2478e+10 1.0443e+12 24138
## - BedroomAbvGr       1 2.2954e+10 1.0448e+12 24138
## - BsmtFullBath        1 2.7198e+10 1.0491e+12 24143
## - OverallCond         1 2.9914e+10 1.0518e+12 24146
## - KitchenAbvGr       1 3.8224e+10 1.0601e+12 24155
## - GarageCars          1 4.5830e+10 1.0677e+12 24164
## - YearBuilt           1 5.8000e+10 1.0799e+12 24177
## - GrLivArea           1 6.9541e+10 1.0914e+12 24189
## - OverallQual         1 1.4122e+11 1.1631e+12 24264
## - gruposHc            1 1.8155e+11 1.2034e+12 24303

# Resumen del modelo2
summary(modelo2)

##
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + OverallCond +
##      YearBuilt + YearRemodAdd + MasVnrArea + TotalBsmtSF + X1stFlrSF +
##      GrLivArea + BsmtFullBath + FullBath + BedroomAbvGr + KitchenAbvGr
+

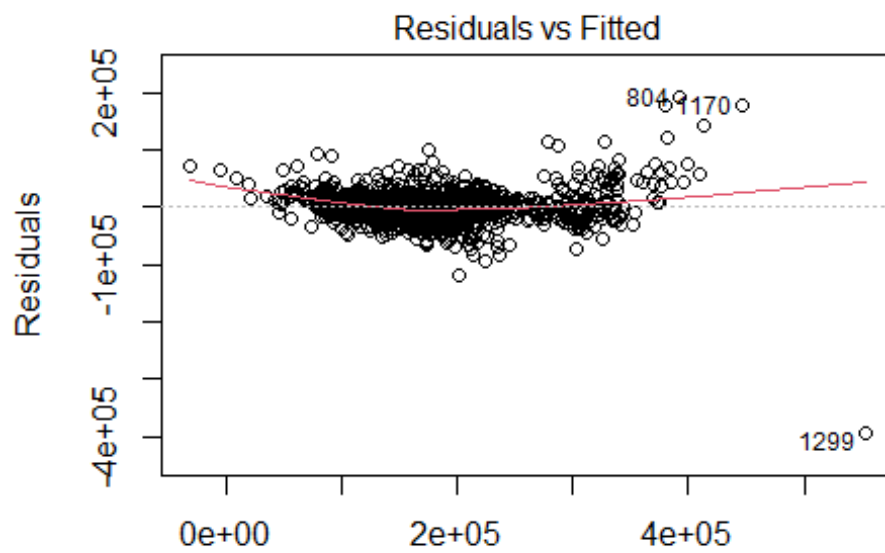
```

```

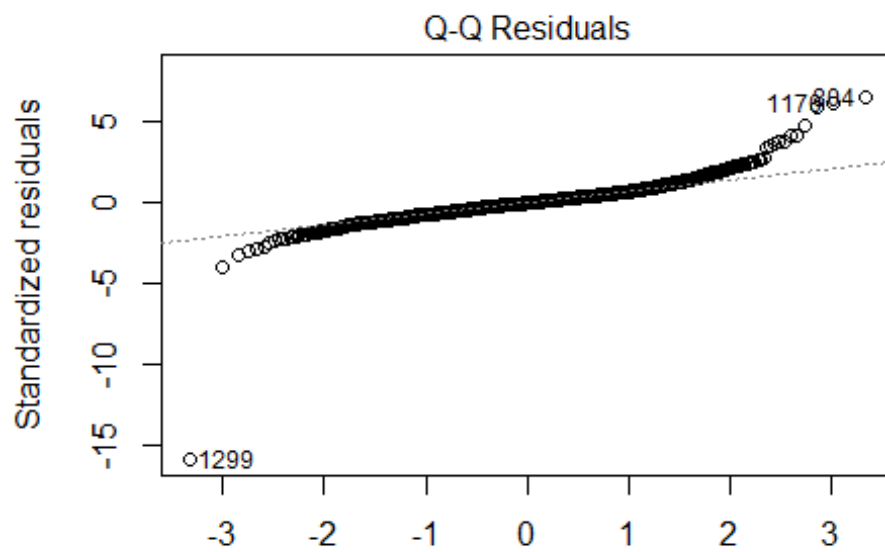
##      TotRmsAbvGrd + Fireplaces + GarageCars + WoodDeckSF + OpenPorchSF
+
##      X3SsnPorch + ScreenPorch + PoolArea + gruposHc, data = train)
##
## Residuals:
##      Min        1Q    Median        3Q        Max
## -392907  -15037   -1071    12929   190852
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.252e+06  1.204e+05 -10.405  < 2e-16 ***
## LotArea      3.753e-01  1.078e-01   3.481 0.000518 ***
## OverallQual  1.410e+04  1.120e+03  12.585  < 2e-16 ***
## OverallCond  5.582e+03  9.637e+02   5.792 8.97e-09 ***
## YearBuilt    4.188e+02  5.193e+01   8.065 1.83e-15 ***
## YearRemodAdd 1.707e+02  6.273e+01   2.721 0.006599 **
## MasVnrArea   1.636e+01  5.798e+00   2.822 0.004860 **
## TotalBsmntSF 1.116e+01  3.833e+00   2.911 0.003672 **
## X1stFlrSF    1.011e+01  4.512e+00   2.241 0.025206 *
## GrLivArea     3.591e+01  4.066e+00   8.831  < 2e-16 ***
## BsmntFullBath 1.038e+04  1.879e+03   5.523 4.12e-08 ***
## FullBath      9.742e+03  2.577e+03   3.781 0.000164 ***
## BedroomAbvGr -8.107e+03  1.598e+03  -5.074 4.55e-07 ***
## KitchenAbvGr -3.026e+04  4.622e+03  -6.547 8.82e-11 ***
## TotRmsAbvGrd  6.020e+03  1.199e+03   5.021 5.96e-07 ***
## Fireplaces    4.533e+03  1.706e+03   2.658 0.007979 **
## GarageCars    1.155e+04  1.611e+03   7.169 1.35e-12 ***
## WoodDeckSF    1.947e+01  7.706e+00   2.527 0.011636 *
## OpenPorchSF   2.924e+01  1.459e+01   2.005 0.045189 *
## X3SsnPorch    5.112e+01  3.214e+01   1.591 0.111952
## ScreenPorch   5.036e+01  1.689e+01   2.981 0.002932 **
## PoolArea     -1.064e+02  2.290e+01  -4.646 3.77e-06 ***
## gruposHc      2.058e+04  1.442e+03  14.269  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29860 on 1146 degrees of freedom
## Multiple R-squared:  0.8493, Adjusted R-squared:  0.8464
## F-statistic: 293.6 on 22 and 1146 DF,  p-value: < 2.2e-16

plot(modelo2)

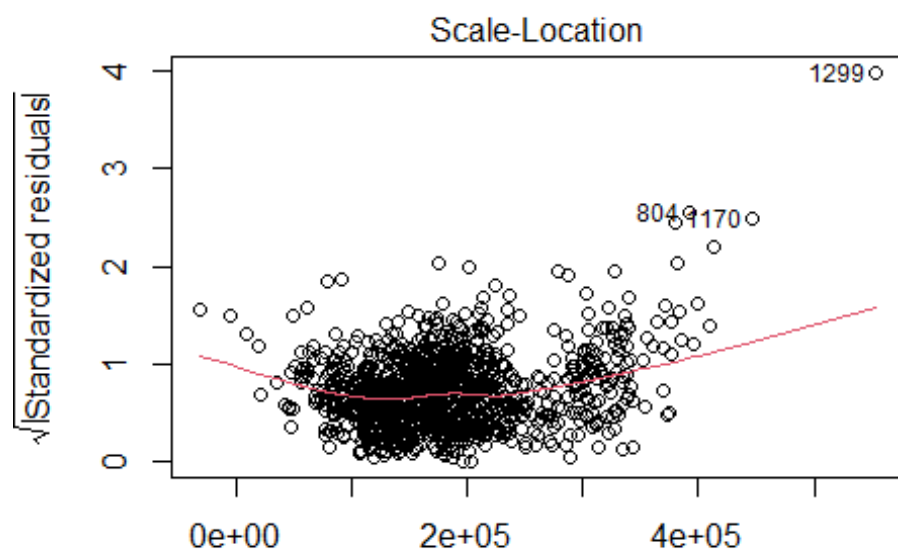
```



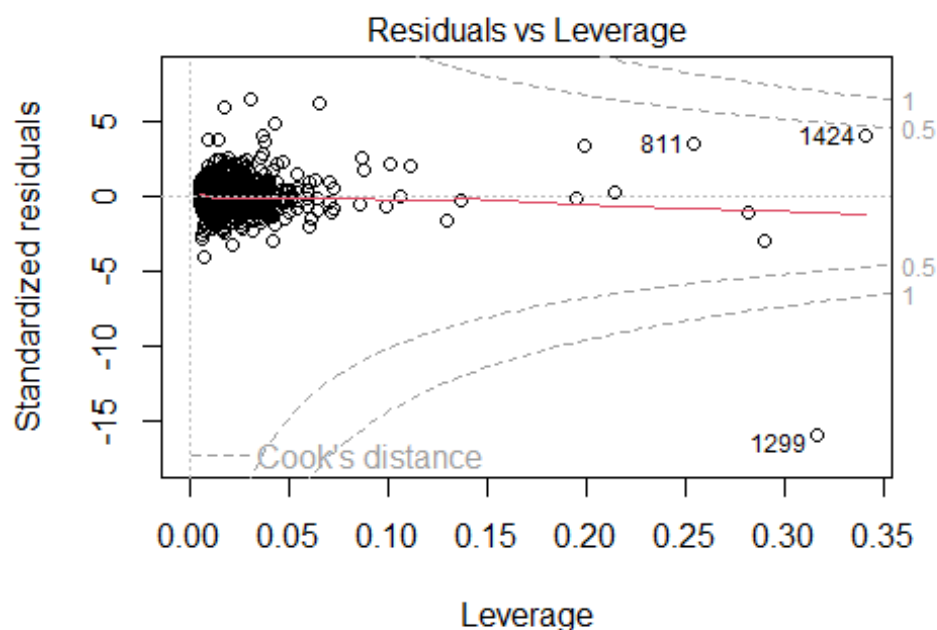
Fitted values
 $\text{SalePrice} \sim \text{LotArea} + \text{OverallQual} + \text{OverallCond} + \text{YearBuilt} + \text{YearRe}$



Theoretical Quantiles
 $\text{SalePrice} \sim \text{LotArea} + \text{OverallQual} + \text{OverallCond} + \text{YearBuilt} + \text{YearRe}$



Fitted values
 $\text{SalePrice} \sim \text{LotArea} + \text{OverallQual} + \text{OverallCond} + \text{YearBuilt} + \text{YearRe}$



Leverage
 $\text{SalePrice} \sim \text{LotArea} + \text{OverallQual} + \text{OverallCond} + \text{YearBuilt} + \text{YearRe}$

Este ultimo modelos aun tiene heterocedasticidad, no parece ser normal, tiene unos 3 puntos atípicos y esplic un 85% de la varianza.

Comparacion de los modelos

```
AIC(modelo0)
## [1] 28414.97
AIC(modelo1)
## [1] 27451.03
AIC(modelo2)
## [1] 27433.72
```

Podemos notar que el AIC del primer modelo es el más alto. Pero entre el modelo1 y el modelo2 el AIC es parecido aunque el modelo2 tiene uno un poco más pequeño. Podemos concluir que el último modelo es el mejor. Dicho eso ningún modelo cumple con los supuestos para que un modelo lineal sea válido. Si analizamos las gráficas podemos notar que el modelo es bueno y cumple mejor los supuestos para los valores de enmedio. Hacer un modelo por cluster podría volver válidos los supuestos.

```
BIC(modelo0)
## [1] 28430.16
BIC(modelo1)
## [1] 27633.34
BIC(modelo2)
## [1] 27555.25
```

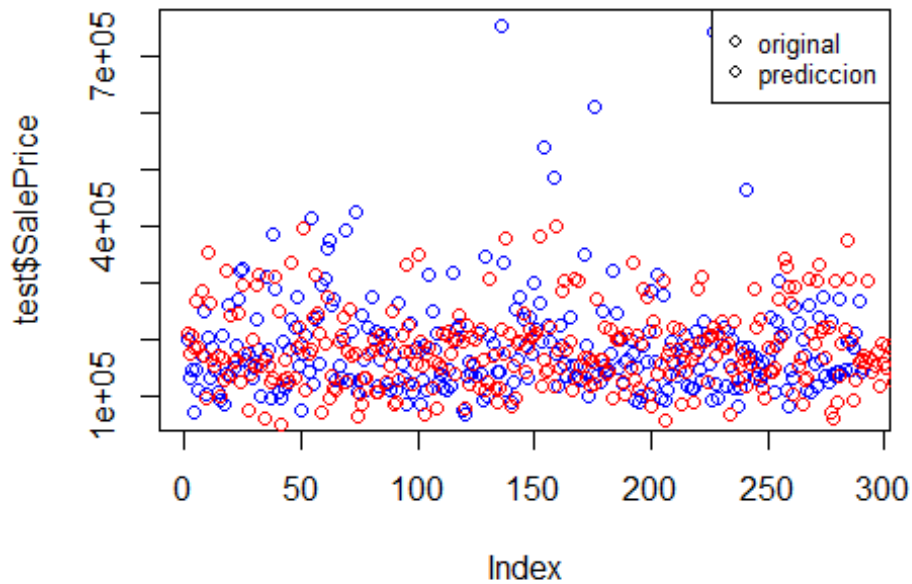
Nótese que el valor más bajo de BIC es del modelo 2, concluyendo que es el mejor, coincidiendo con AIC.

Veamos que tan bueno es el modelo, para ello la gráfica siguiente de valores reales y las predicciones para SalePrice.

```
pred_modelo2 <- predict(modelo2, newdata = train)

#grafico
plot(test$SalePrice, col="blue", main = "prediccion de Modelo 2 vs valores originales")
points(pred_modelo2, col="red")
legend("topright", legend = c("original", "prediccion"), col = c("blue", "red"), pch=1, cex=0.8)
```


prediccion de Modelo 2 vs valores originales



Los puntos azules representan los datos reales de SalePrice en el conjunto de prueba, los puntos rojos las predicciones segun el modelo 2 para SalePrice. Nótese que en general los puntos rojos, que son las predicciones siguen las tendencias de los puntos azules, valores reales. Indicando que el modelo capta las variaciones, aunque hay diferencia significativa entre la prediccion y el valor real en algunos puntos. Por lo que el modelo tiene capacidad de captruara algunas tendencias pero tiene limitaciones.