

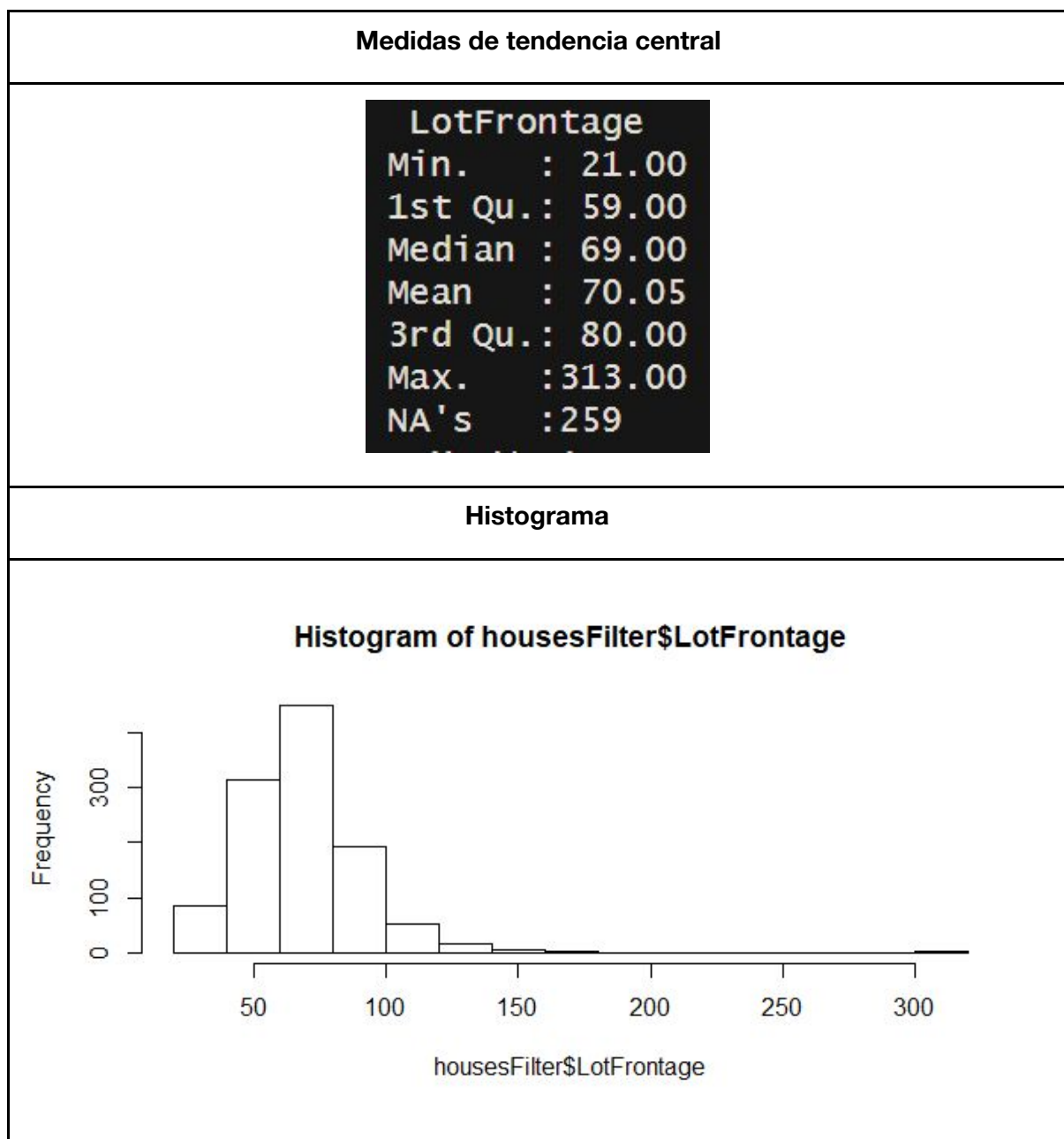
## Hoja de trabajo 3. Árboles de decisión

### 1. Análisis exploratorio:

#### Variables filtradas:

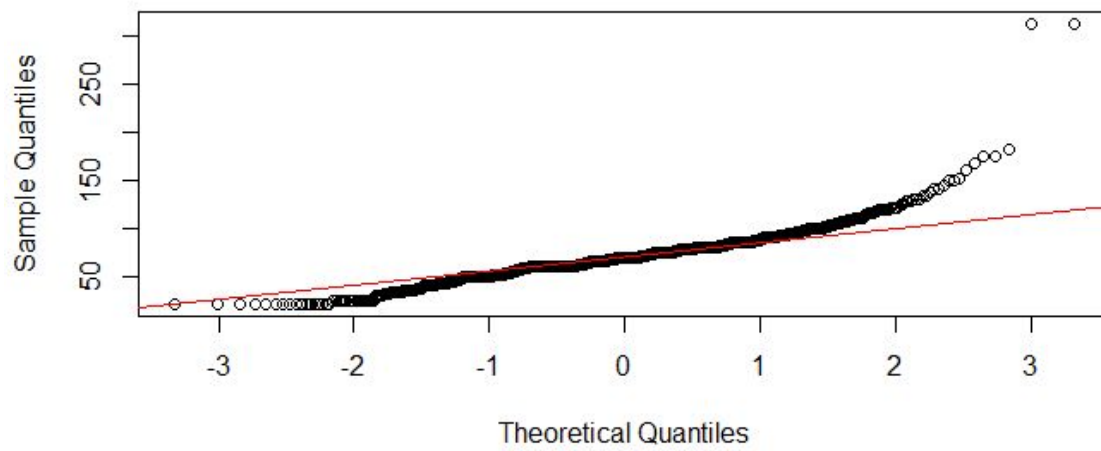
- **LotFrontage:** Metros lineales de la calle conectados a la propiedad.

La variable LotFrontage tiene un sesgo positivo, los datos están ligeramente normalizados y no tienen relación directa con el precio de venta.

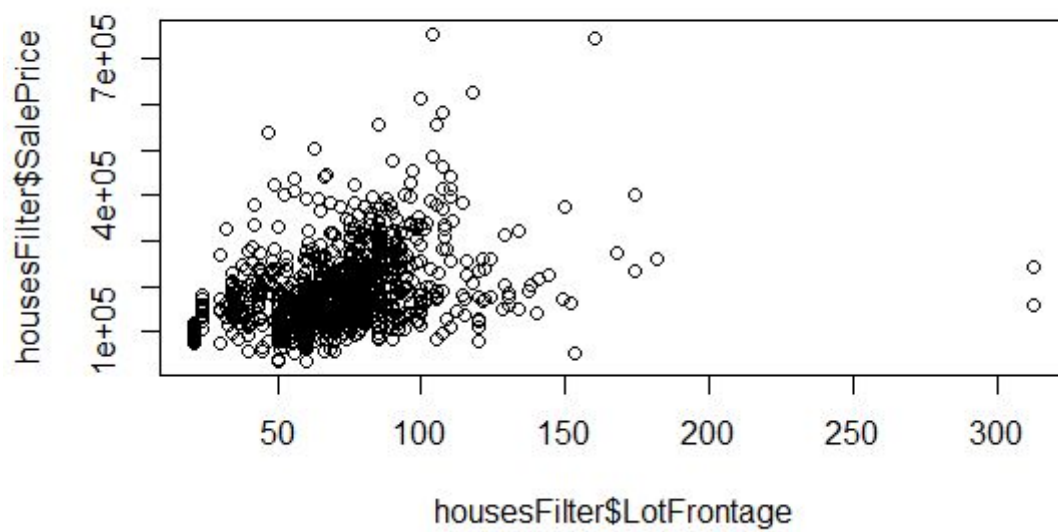


**Gráfica Q-Q**

**Normal Q-Q Plot**

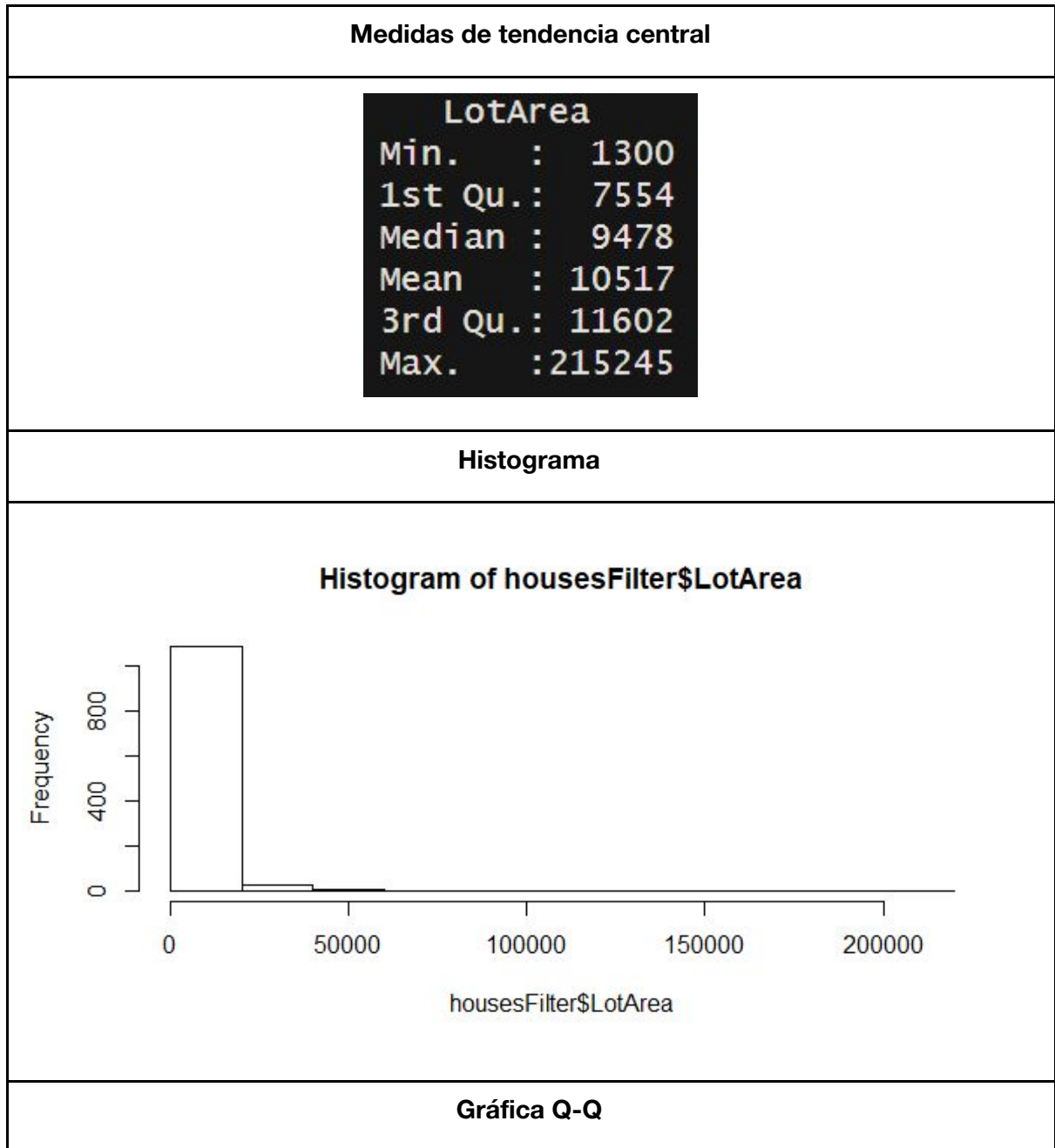


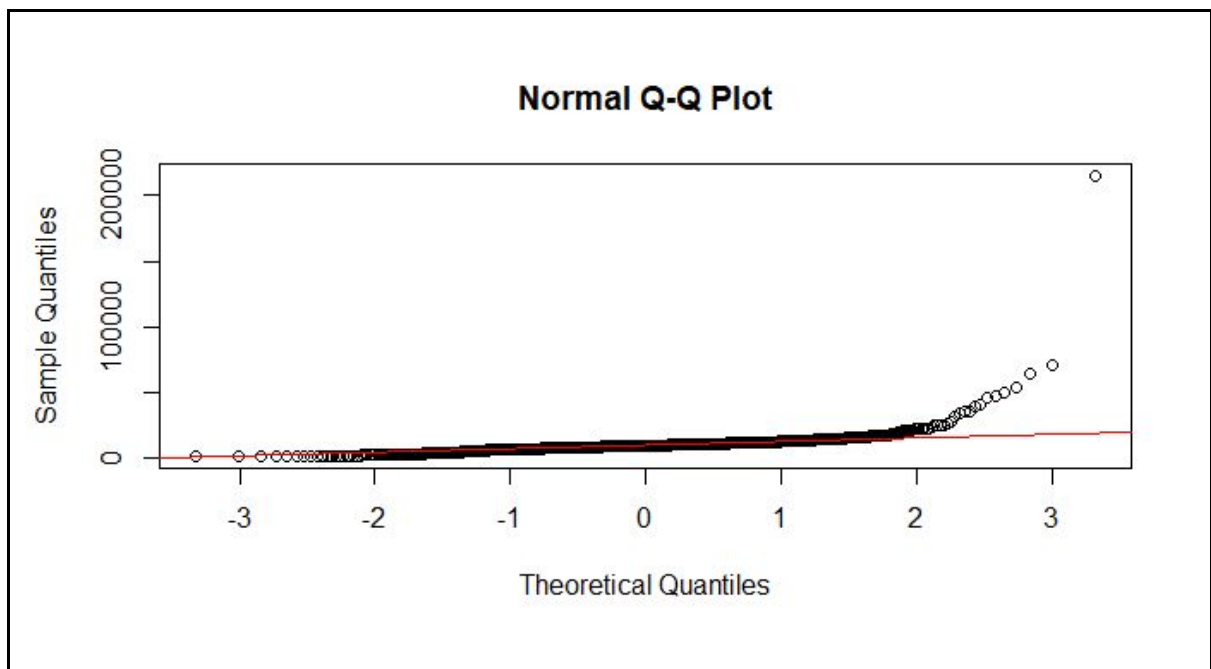
**Gráfica contra precio de venta**



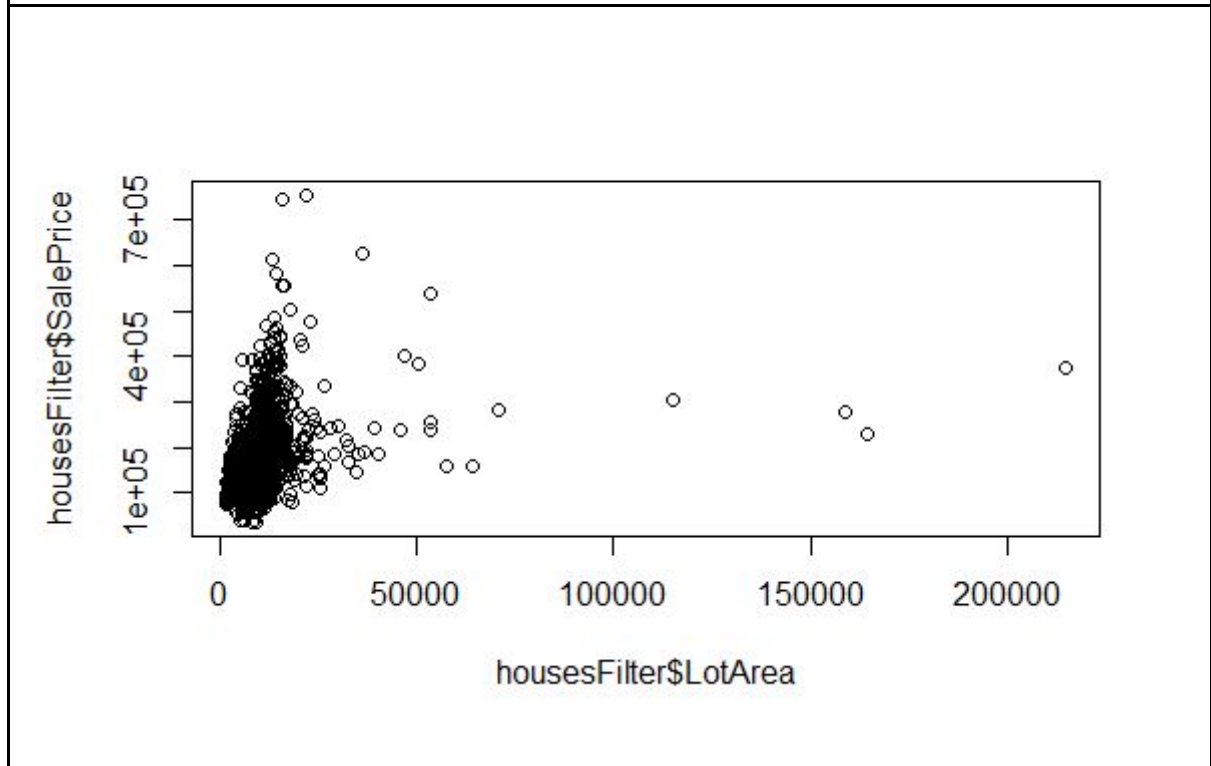
- **LotArea:** Tamaño del lote en metros cuadrados.

La variable LotArea tiene un sesgo positivo, los datos no están muy normalizados y no tienen relación directa con el precio de venta.





Gráfica contra precio de venta



- **YearBuilt:** Año de construcción.

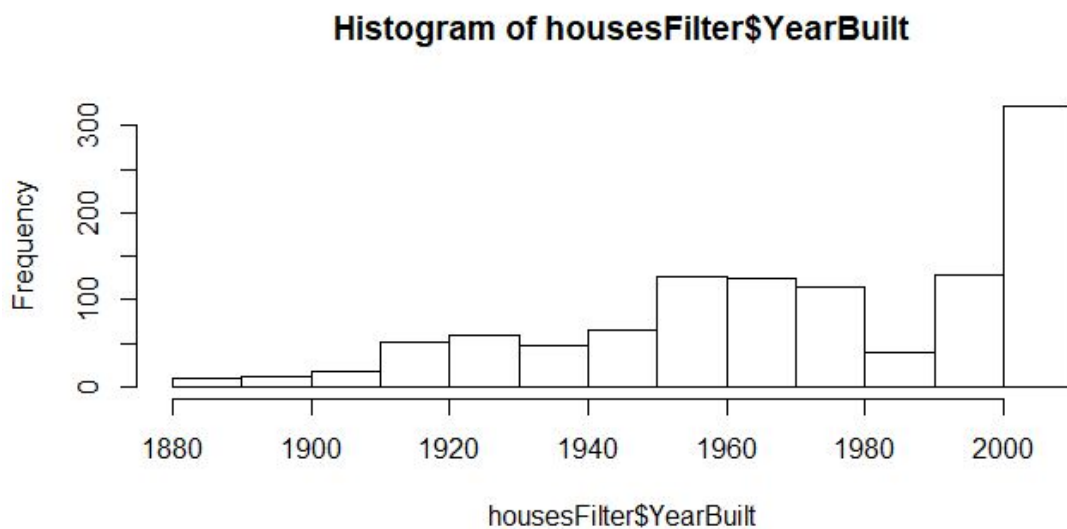
La variable YearBuilt tiene un sesgo negativo, los datos están normalizados en una porción de datos y se puede decir que los últimos años ha aumentado el precio de venta.

### Medidas de tendencia central

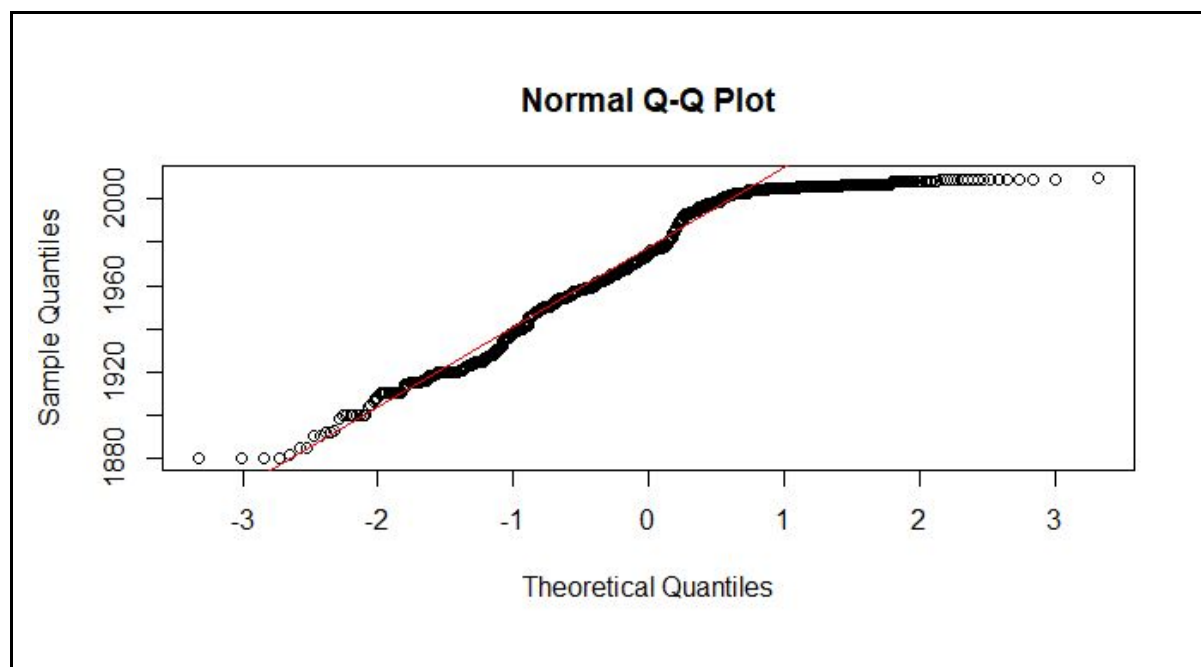
```

YearBuilt
Min.      :1872
1st Qu.   :1954
Median    :1973
Mean      :1971
3rd Qu.   :2000
Max.      :2010
  
```

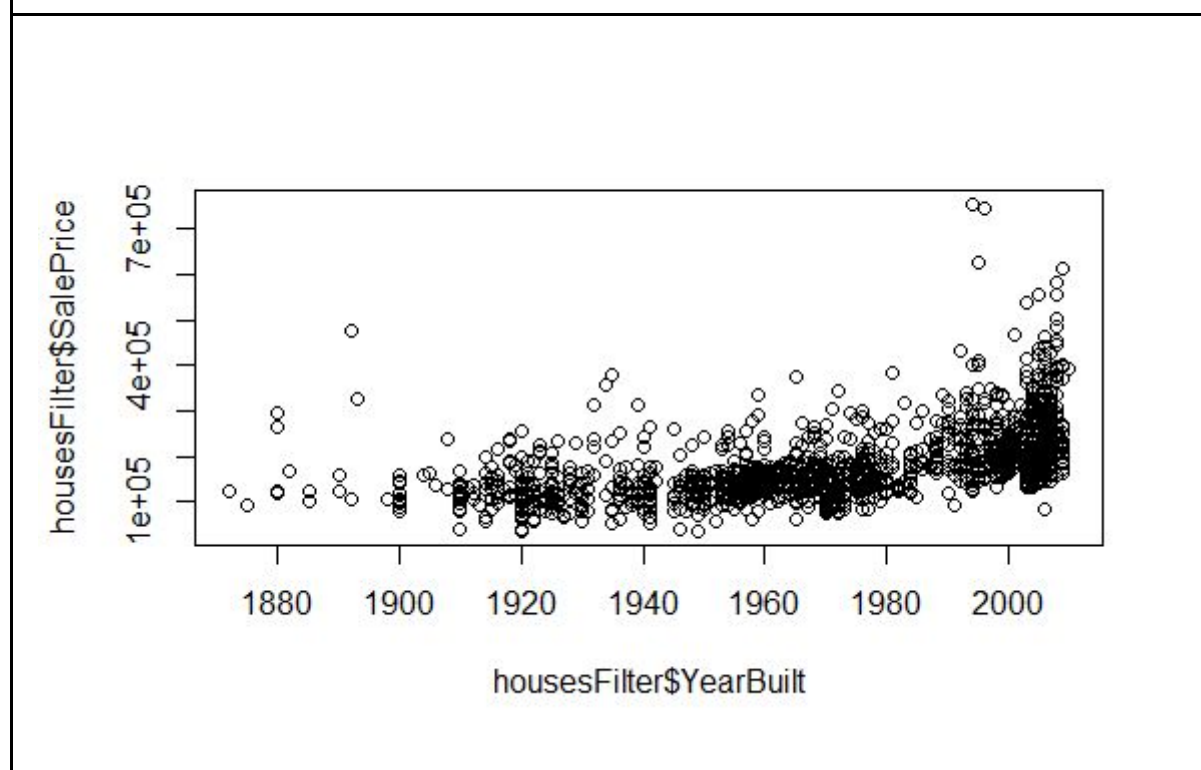
### Histograma



### Gráfica Q-Q

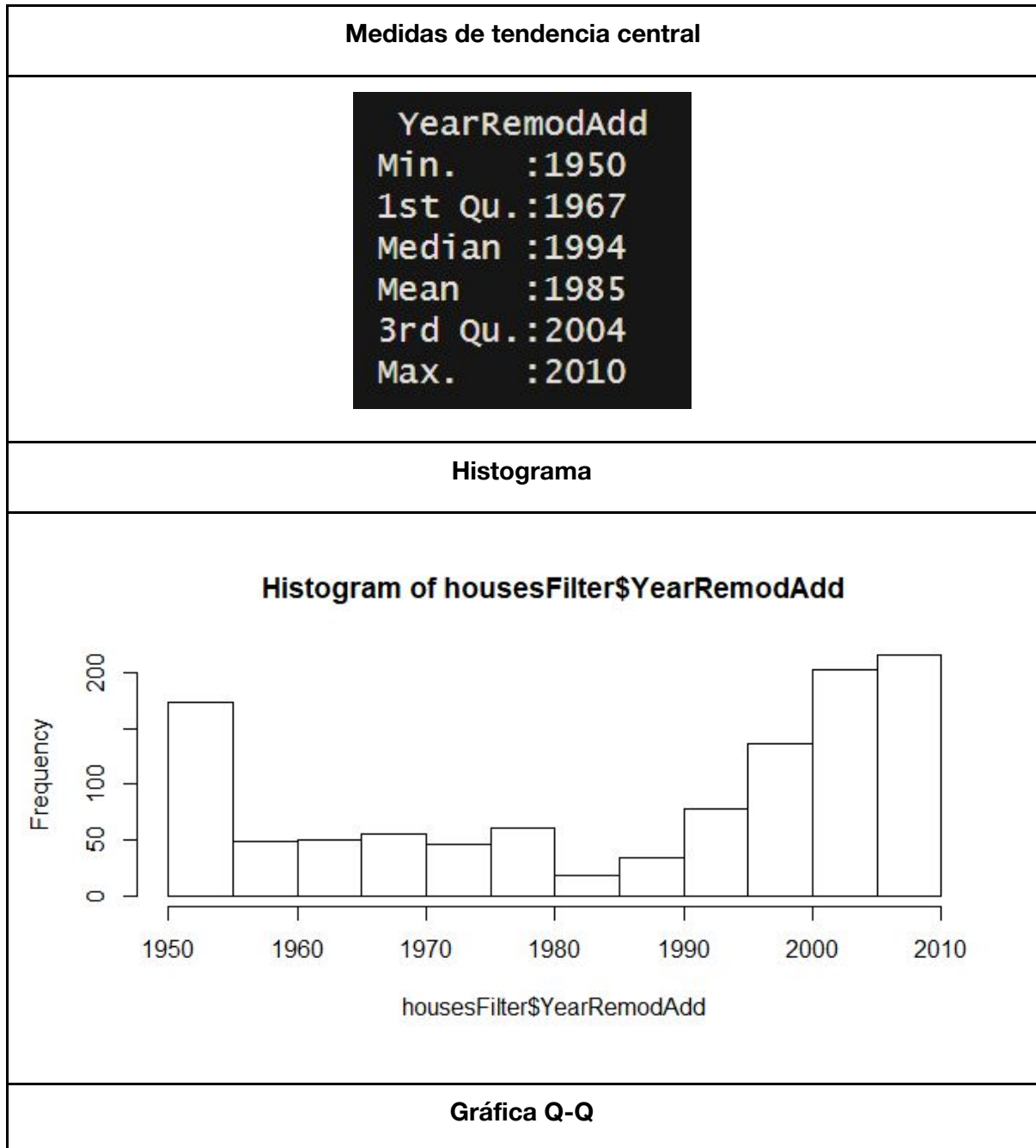


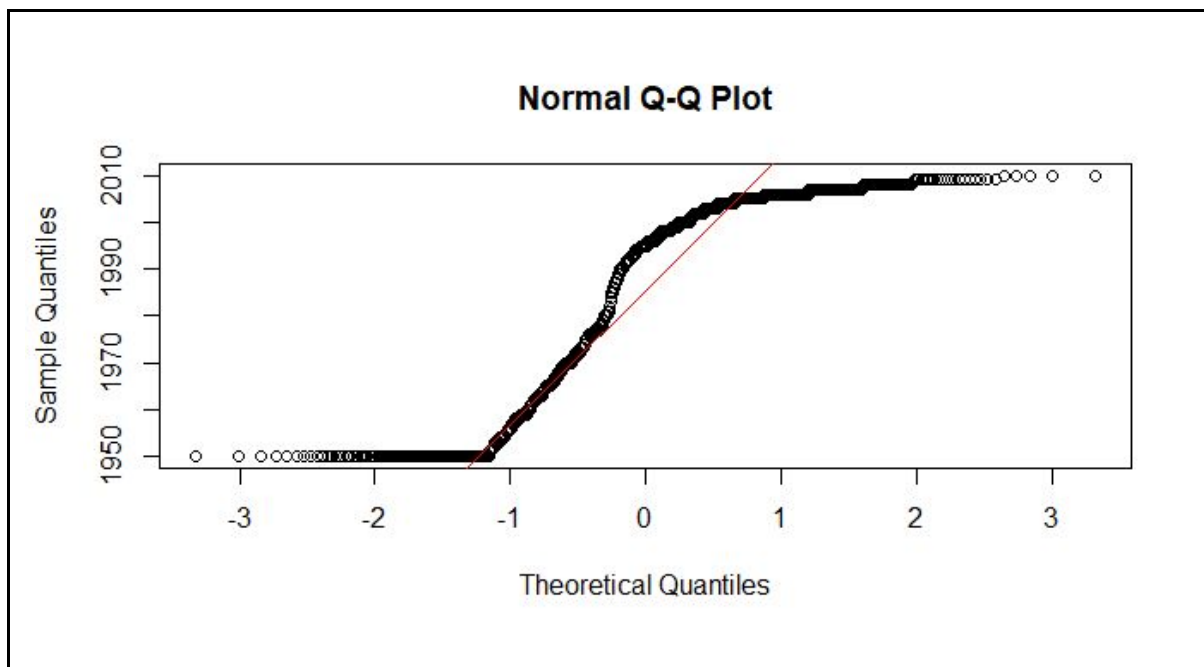
Gráfica contra precio de venta



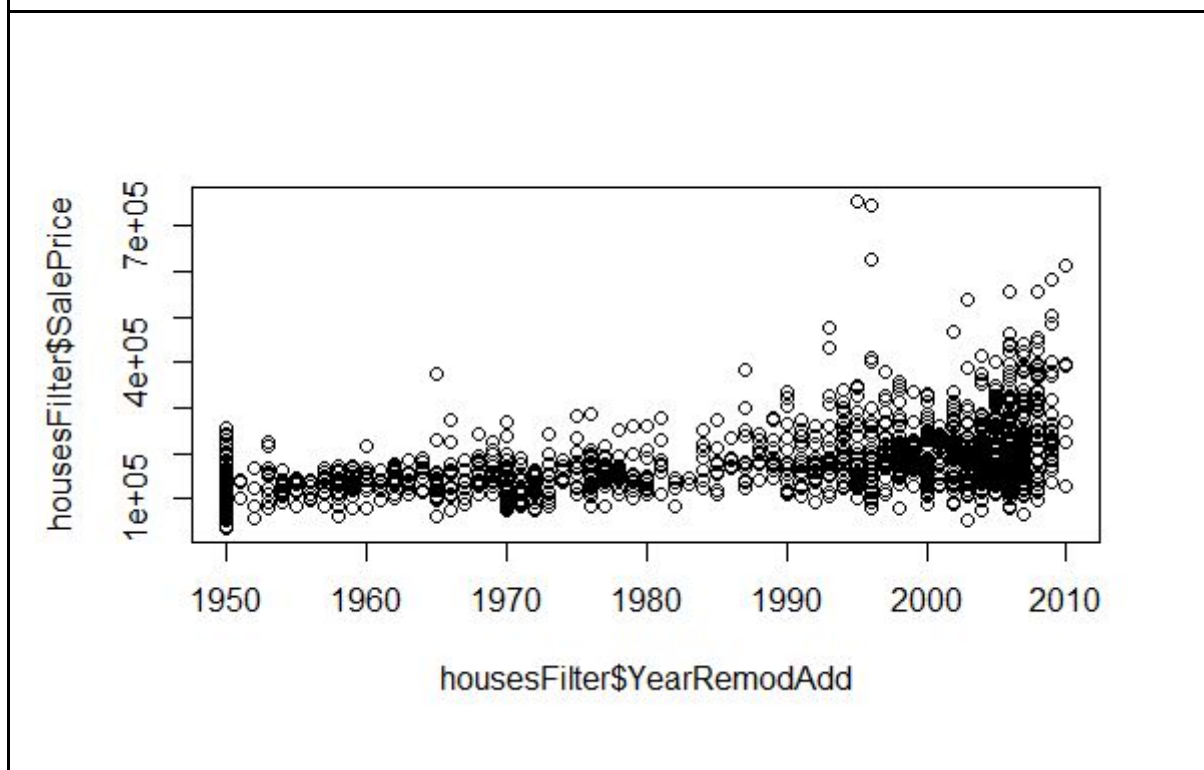
- **YearRemodAdd:** Año de remodelación.

La variable YearRemodAdd no muestra datos significantes en el histograma, los datos no están normalizados y no tienen relación directa con el precio de venta.





Gráfica contra precio de venta





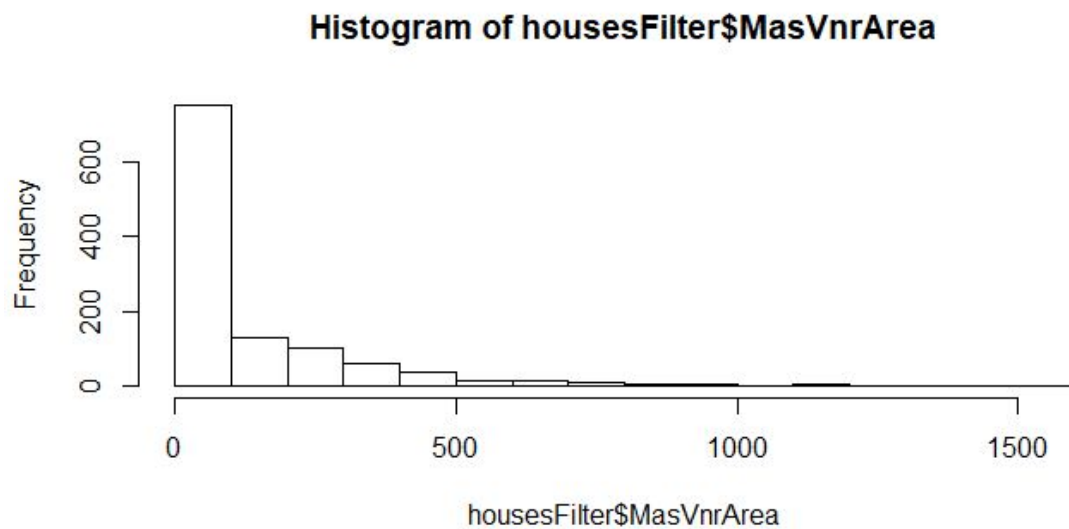
- **MasVnrArea:** Área de mampostería en metros cuadrados.

La variable MasVnrArea tiene un sesgo positivo, los datos no están normalizados y no tienen relación directa con el precio de venta.

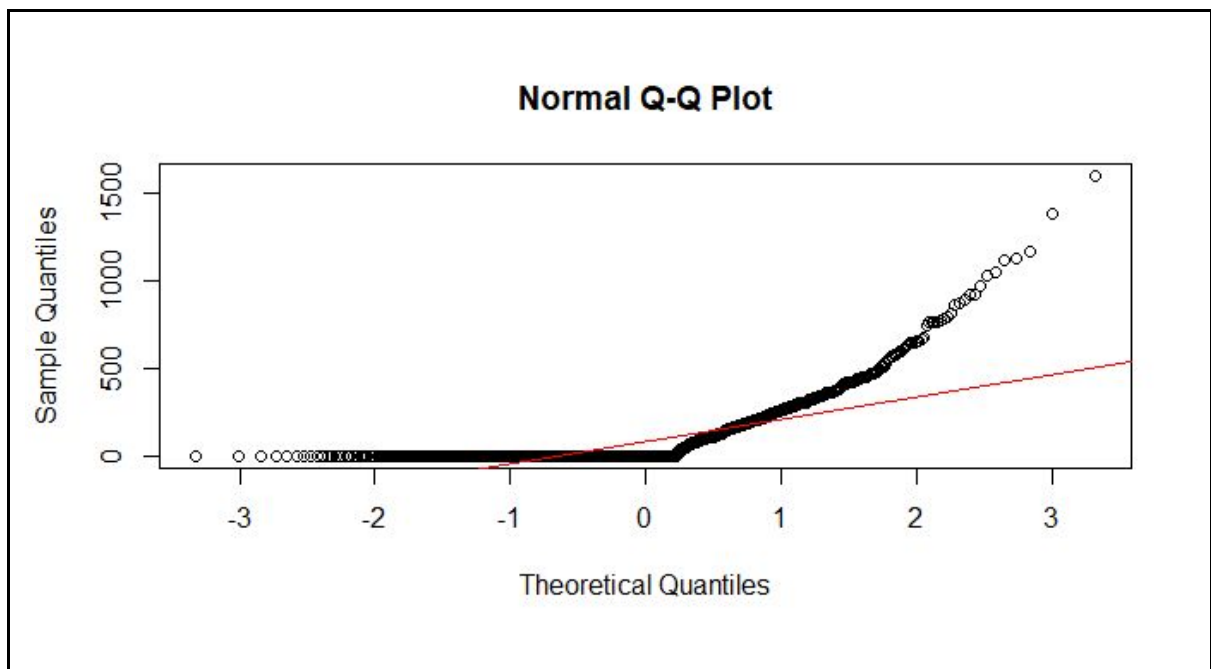
#### Medidas de tendencia central

```
MasVnrArea
Min.   : 0.0
1st Qu.: 0.0
Median : 0.0
Mean   : 103.7
3rd Qu.: 166.0
Max.   : 1600.0
NA's   : 8
```

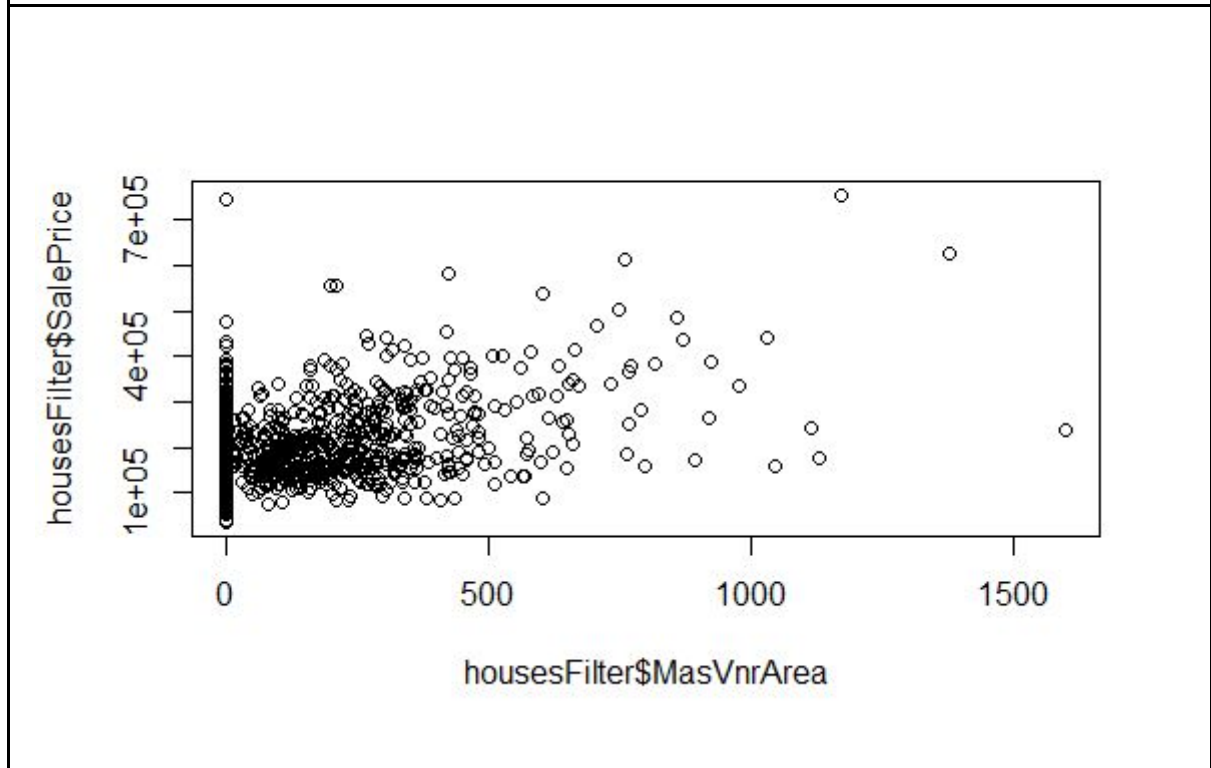
#### Histograma



#### Gráfica Q-Q

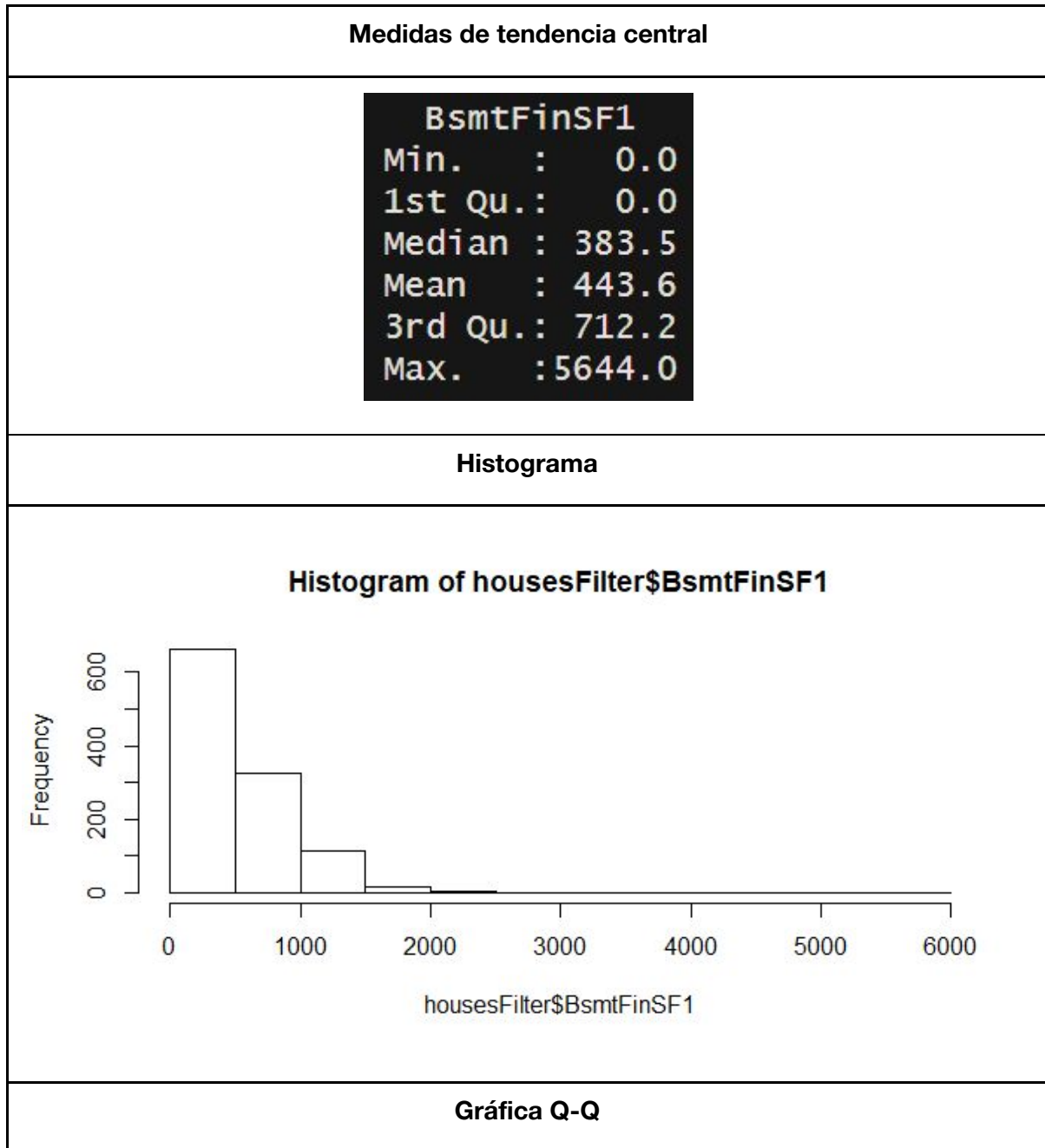


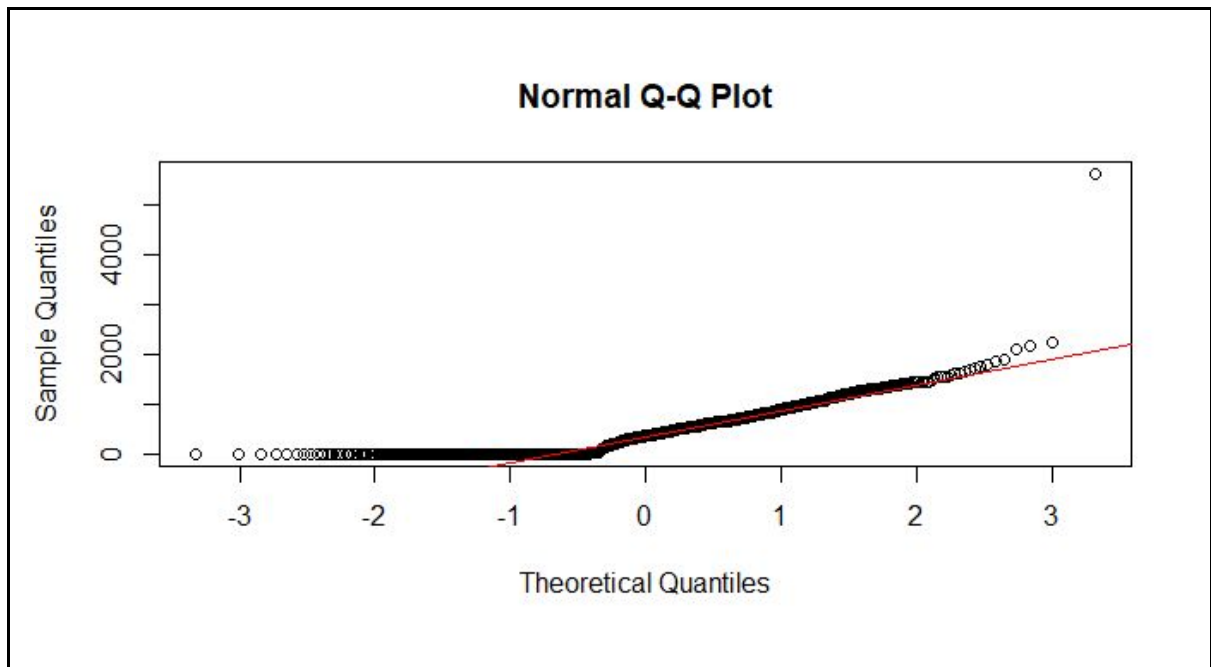
Gráfica contra precio de venta



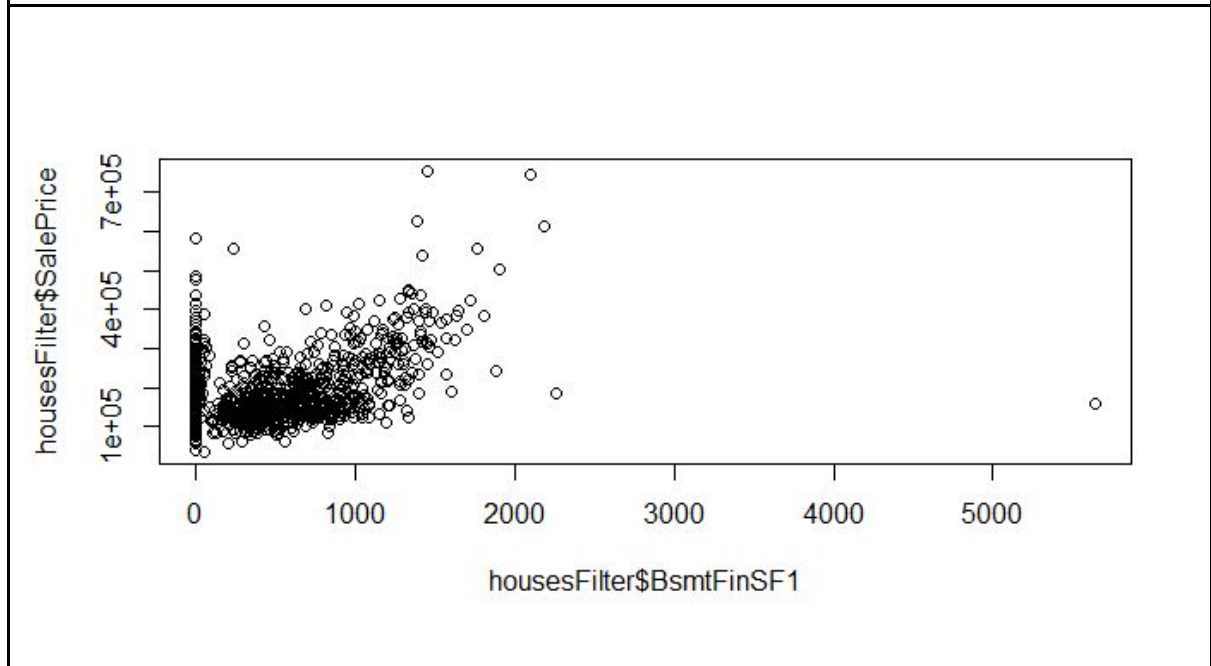
- **BsmtFinSF1:** Metros cuadrados terminados tipo 1.

La variable BsmtFinSF1 tiene un sesgo positivo, los datos sí están normalizados si se ignoran los 0's del dataset y no tienen relación directa con el precio de venta.



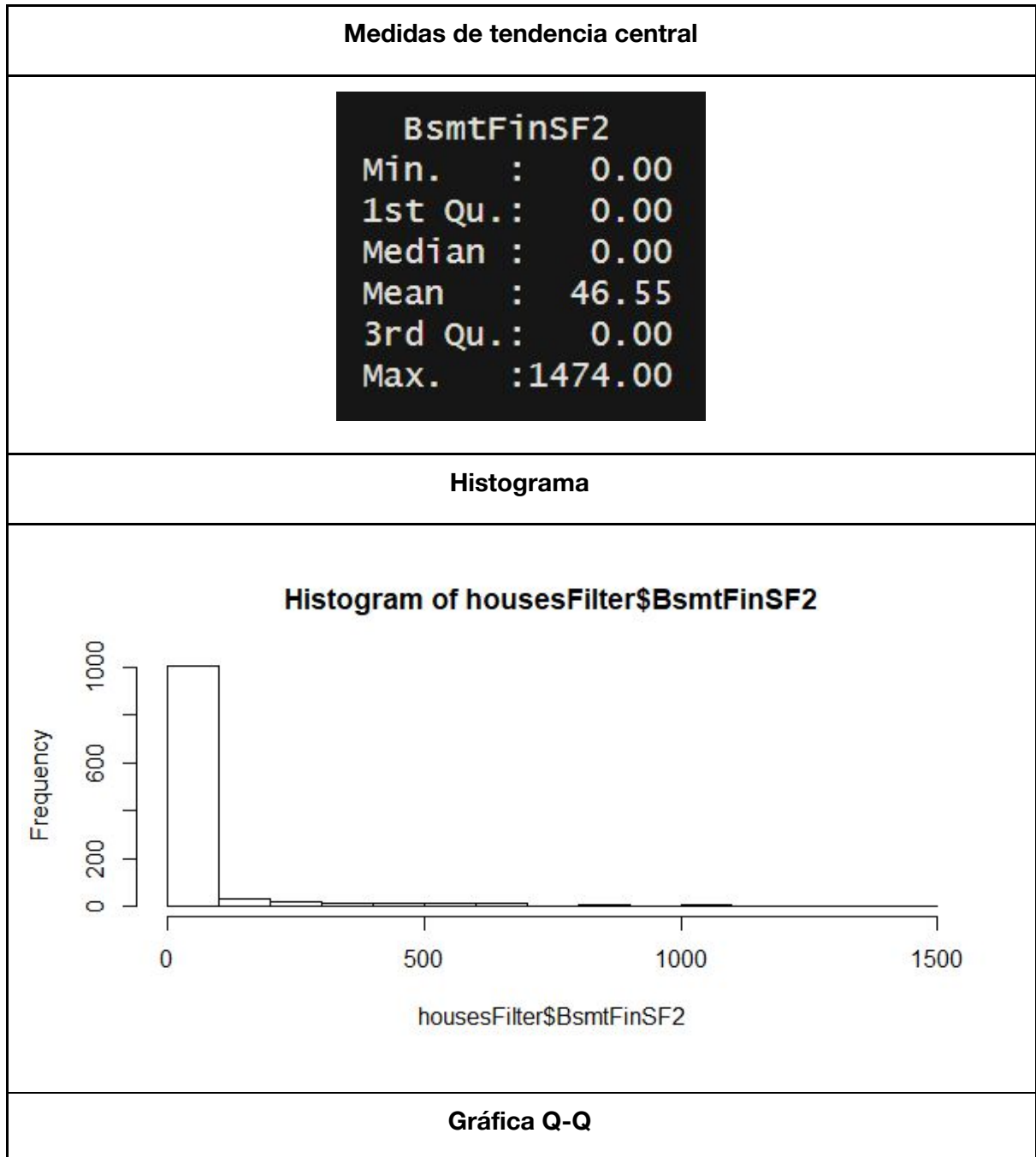


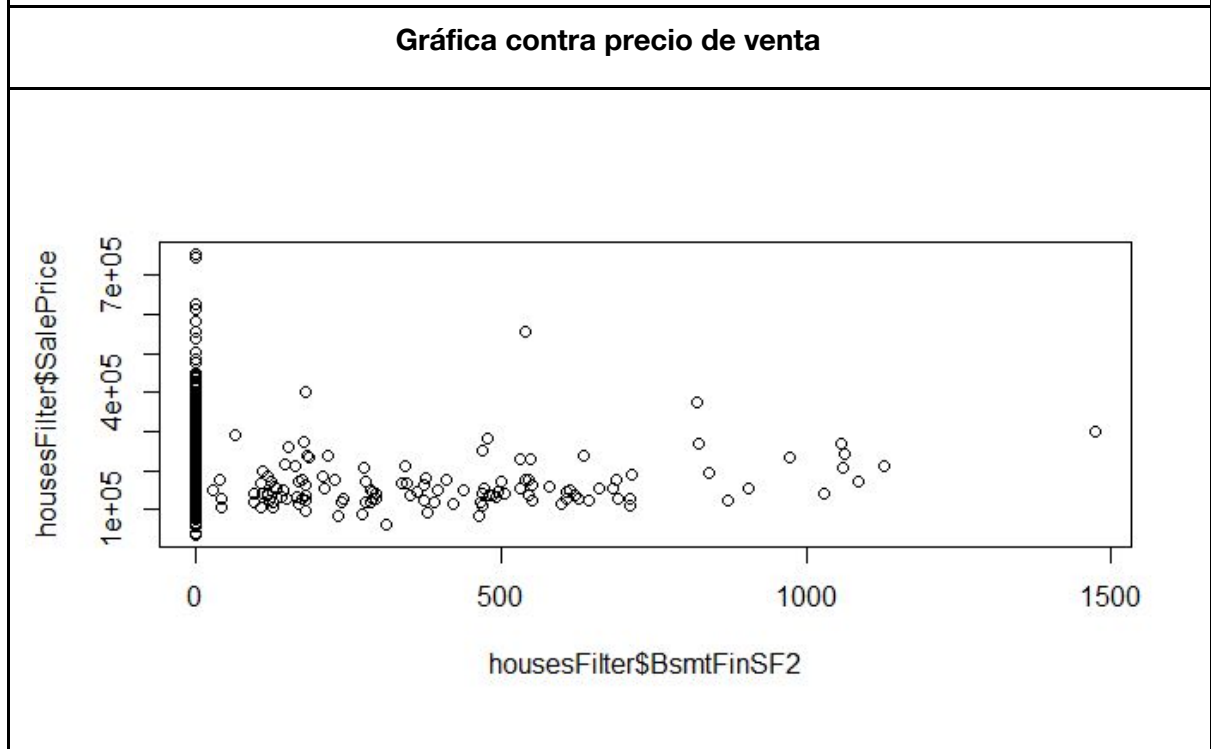
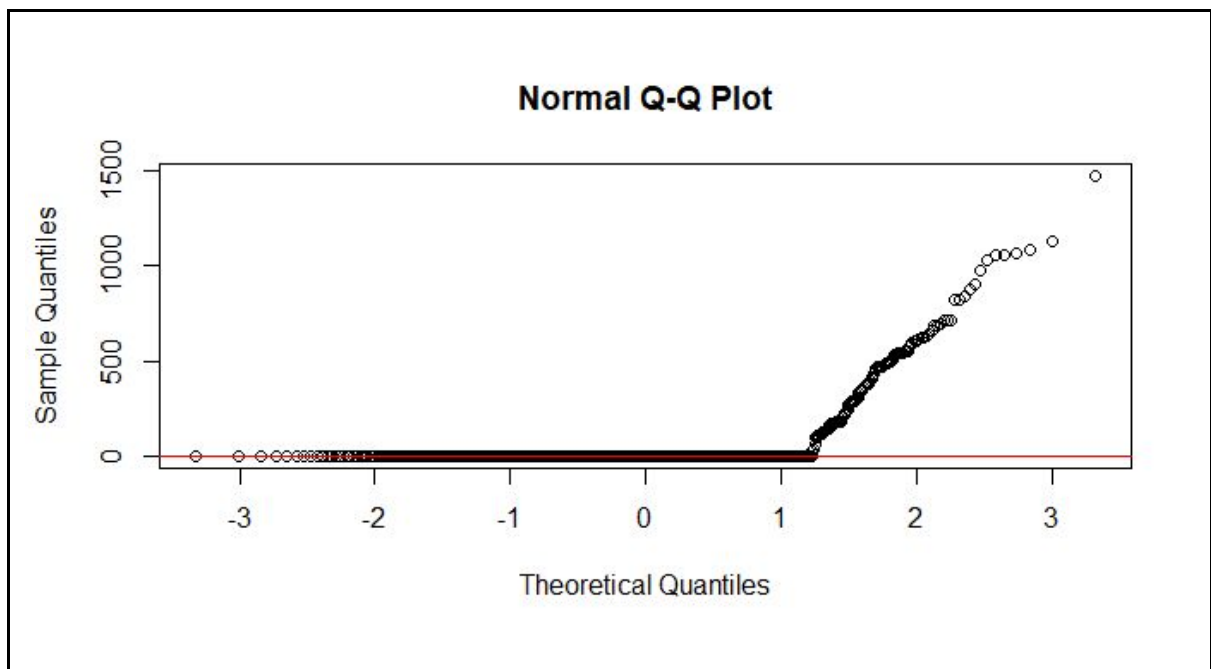
Gráfica contra precio de venta



- **BsmtFinSF2:** Metros cuadrados terminados tipo 2.

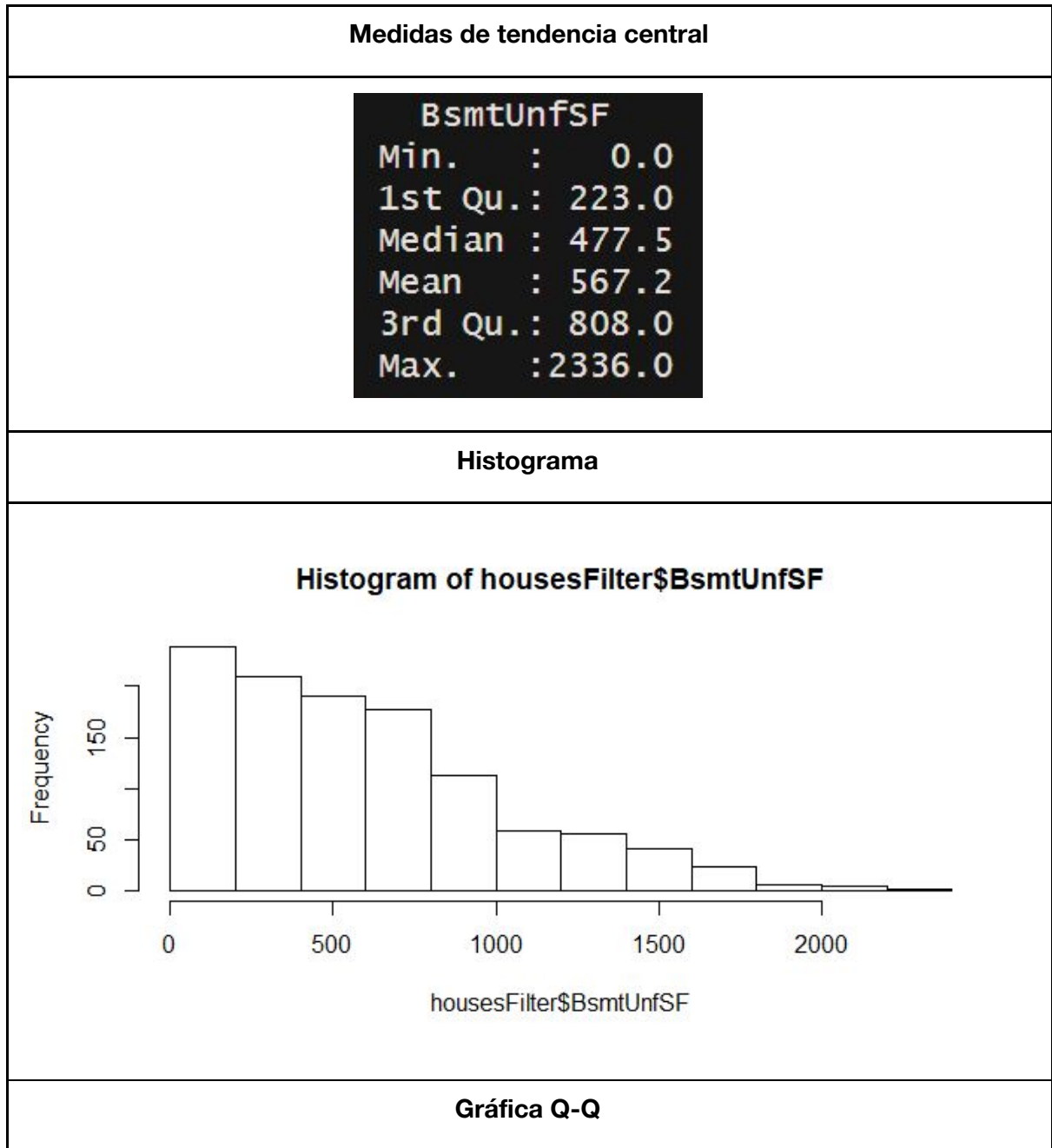
La variable BsmtFinSF2 tiene un sesgo positivo, los datos no están normalizados y no tienen relación directa con el precio de venta.

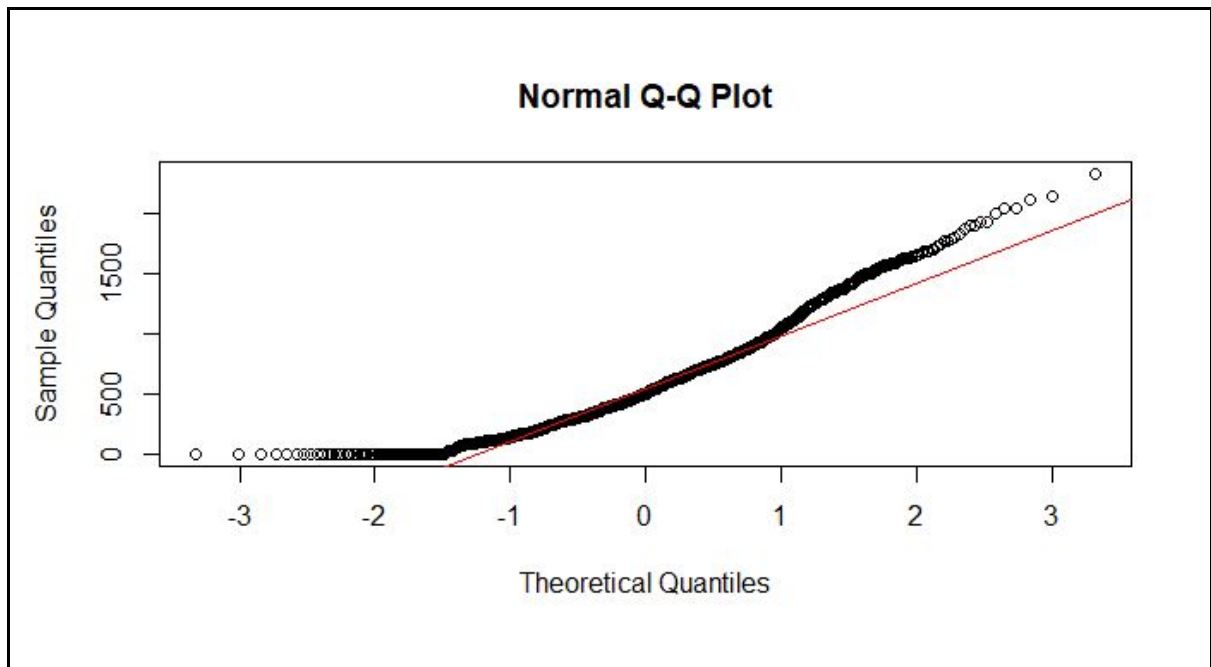




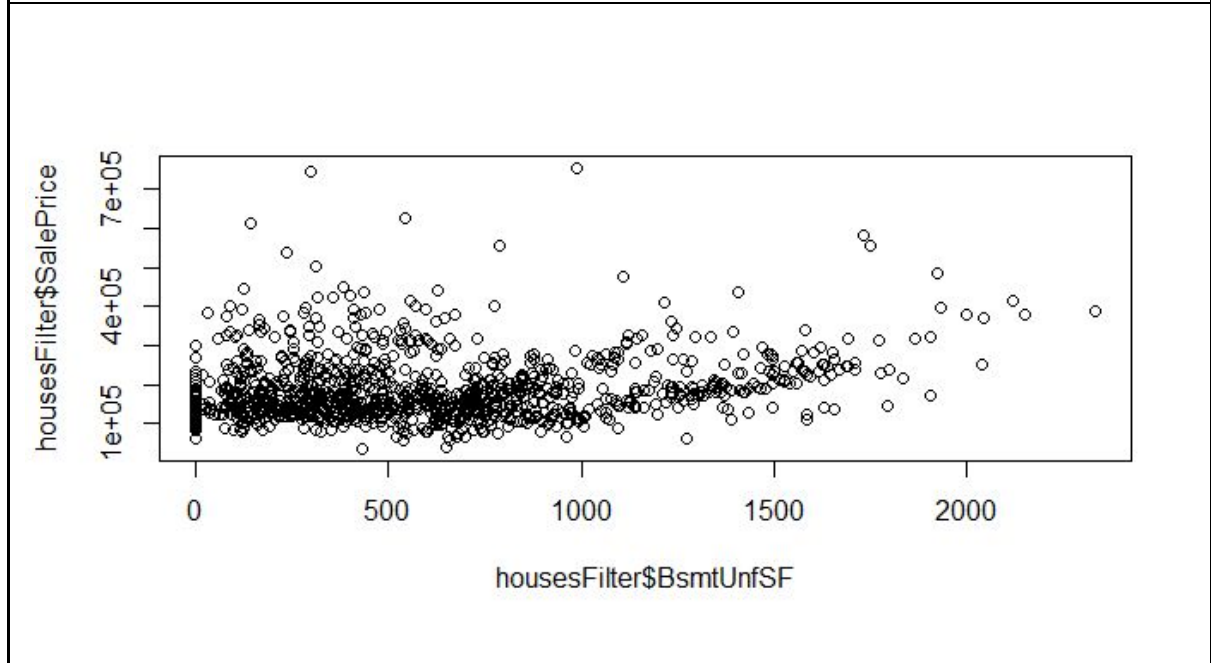
- **BsmtUnfSF:** Metros cuadrados sin terminar en el sótano.

La variable BsmtUnfSF tiene un sesgo positivo, los datos sí están normalizados y tiene relación muy pequeña con el precio de venta.





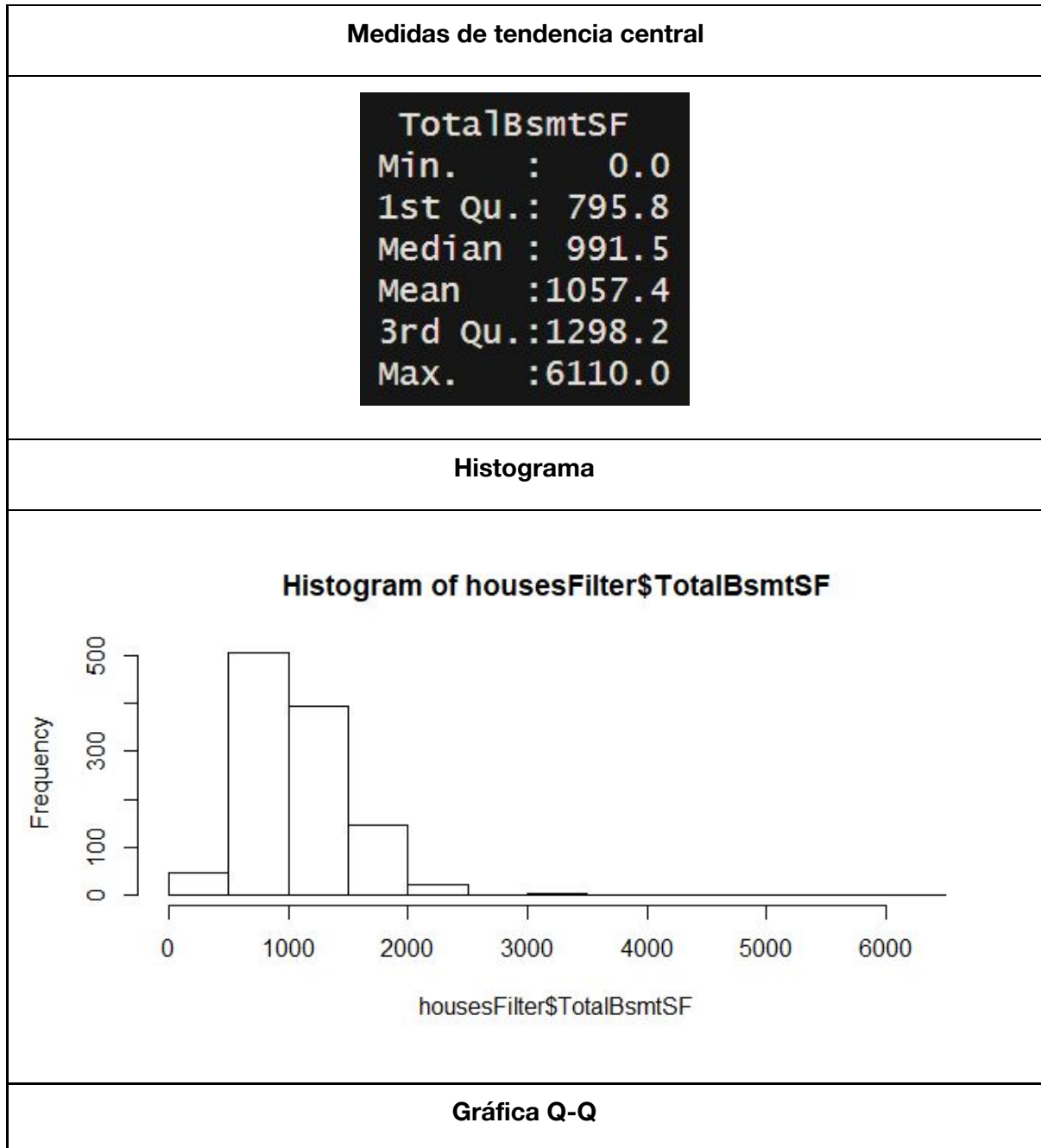
**Gráfica contra precio de venta**

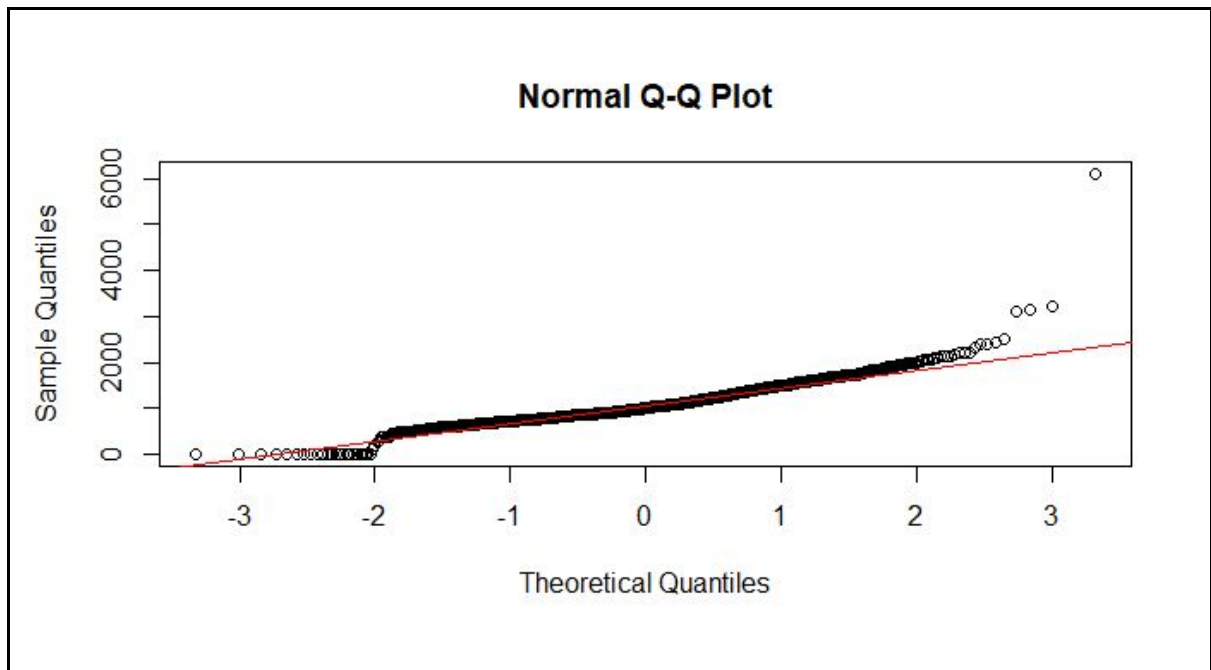




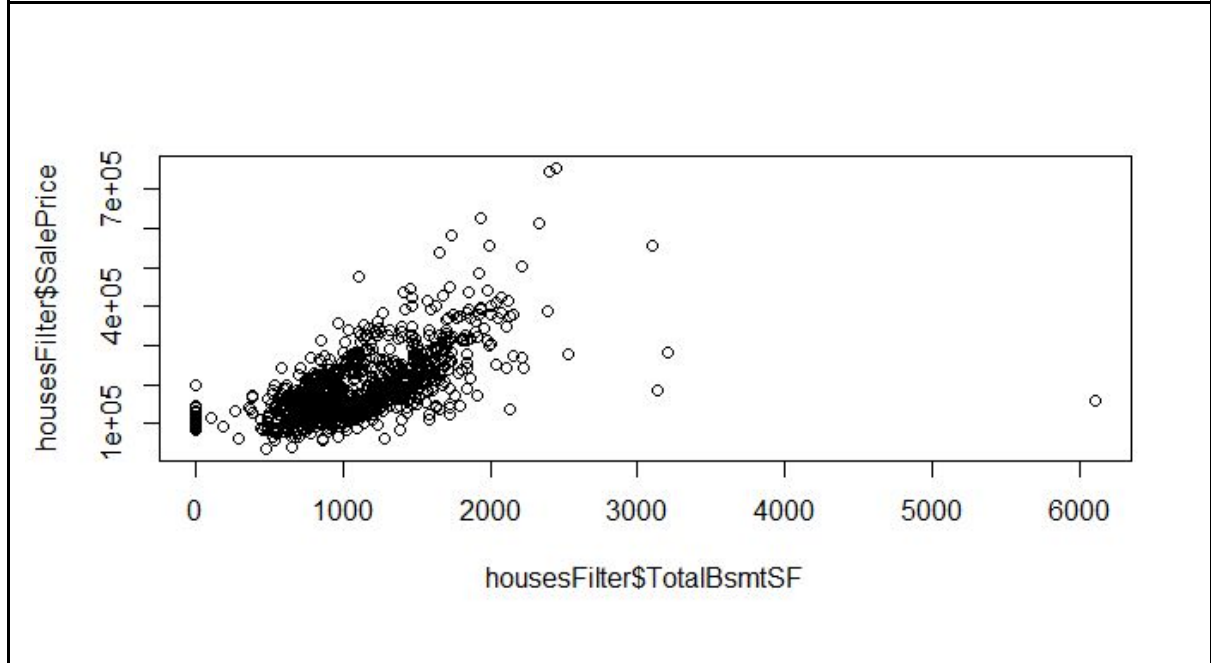
- **TotalBsmtSF:** Metros cuadrados totales del sótano.

La variable TotalBsmtSF tiene un poco de simetría, los datos sí están normalizados y sí tienen relación directa con el precio de venta.



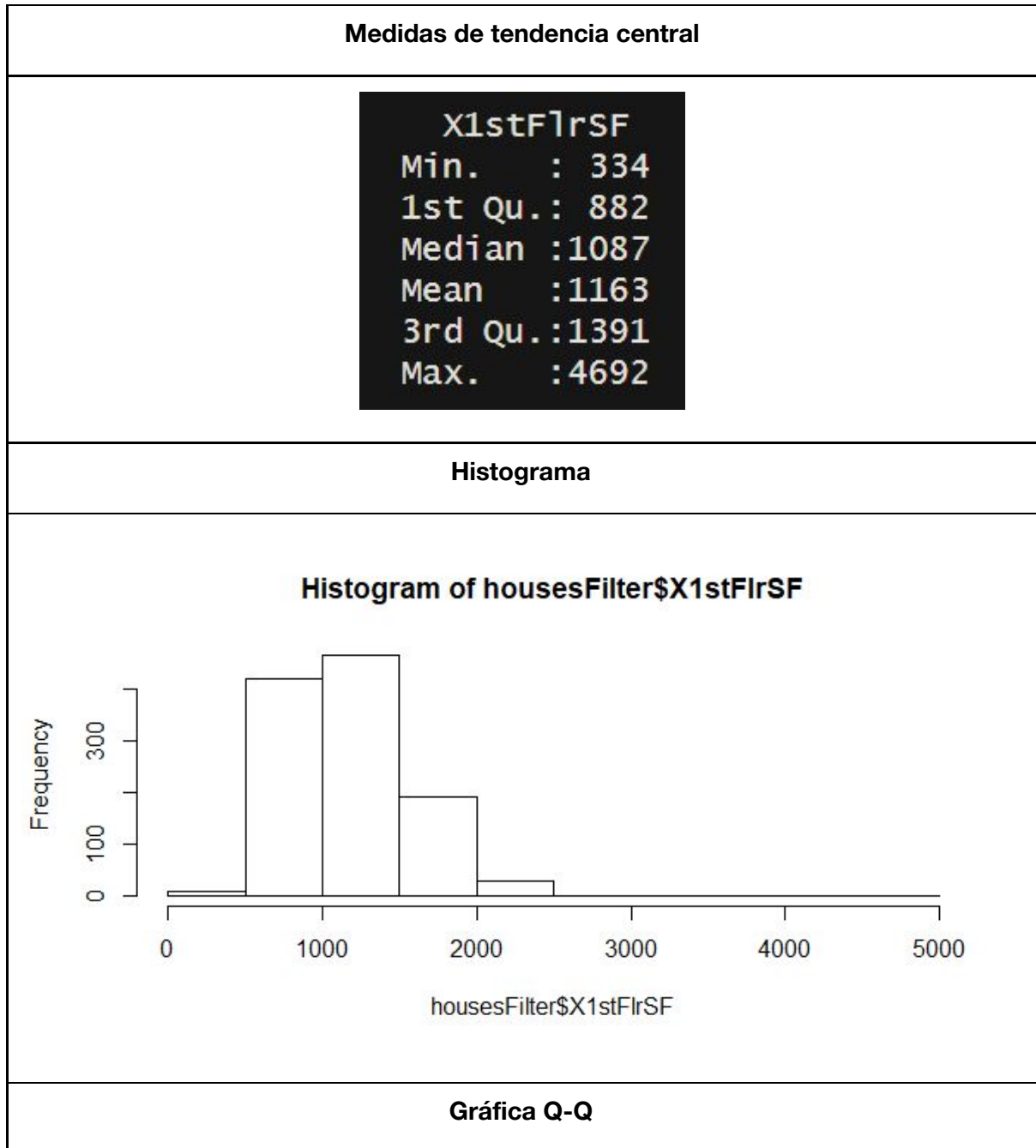


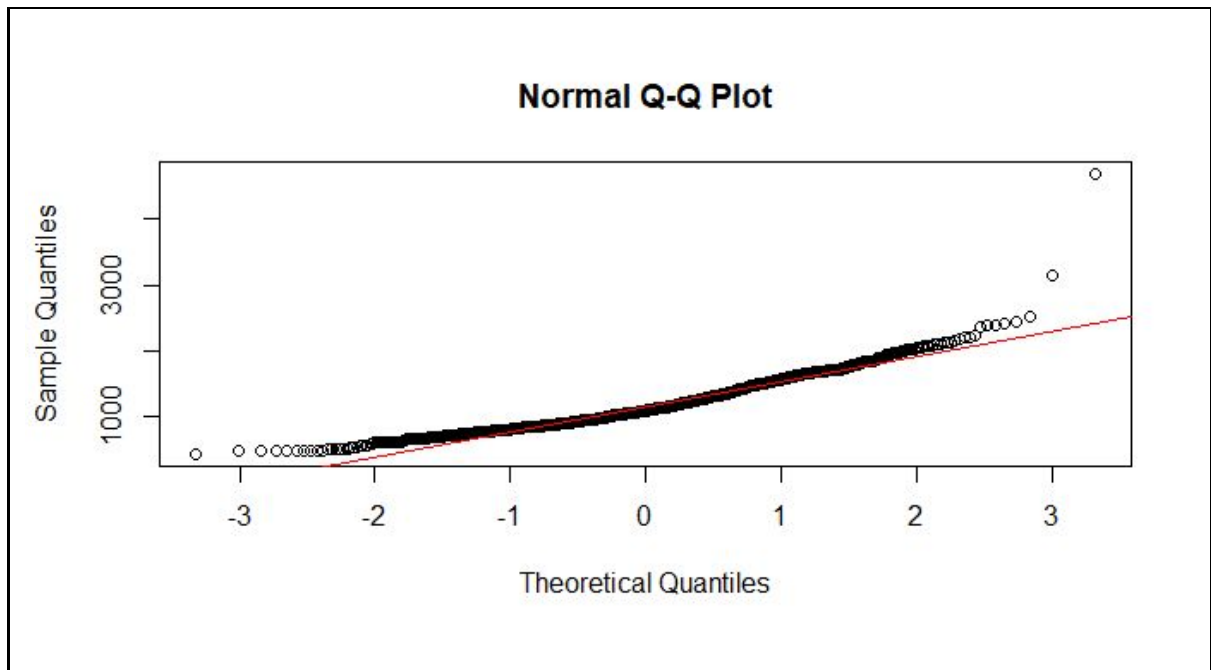
Gráfica contra precio de venta



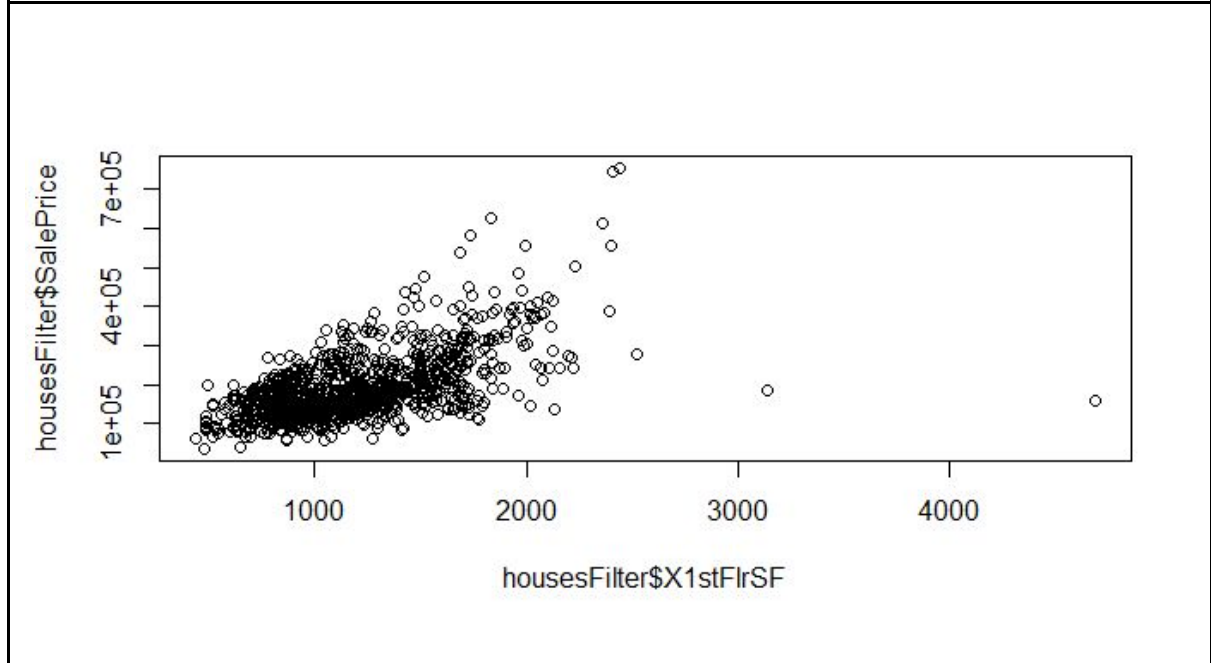
- **X1stFlrSF:** Metros cuadrados del primer piso.

La variable X1stFlrSF tiene un poco de simetría, los datos están normalizados y tienen relación directa con el precio de venta.



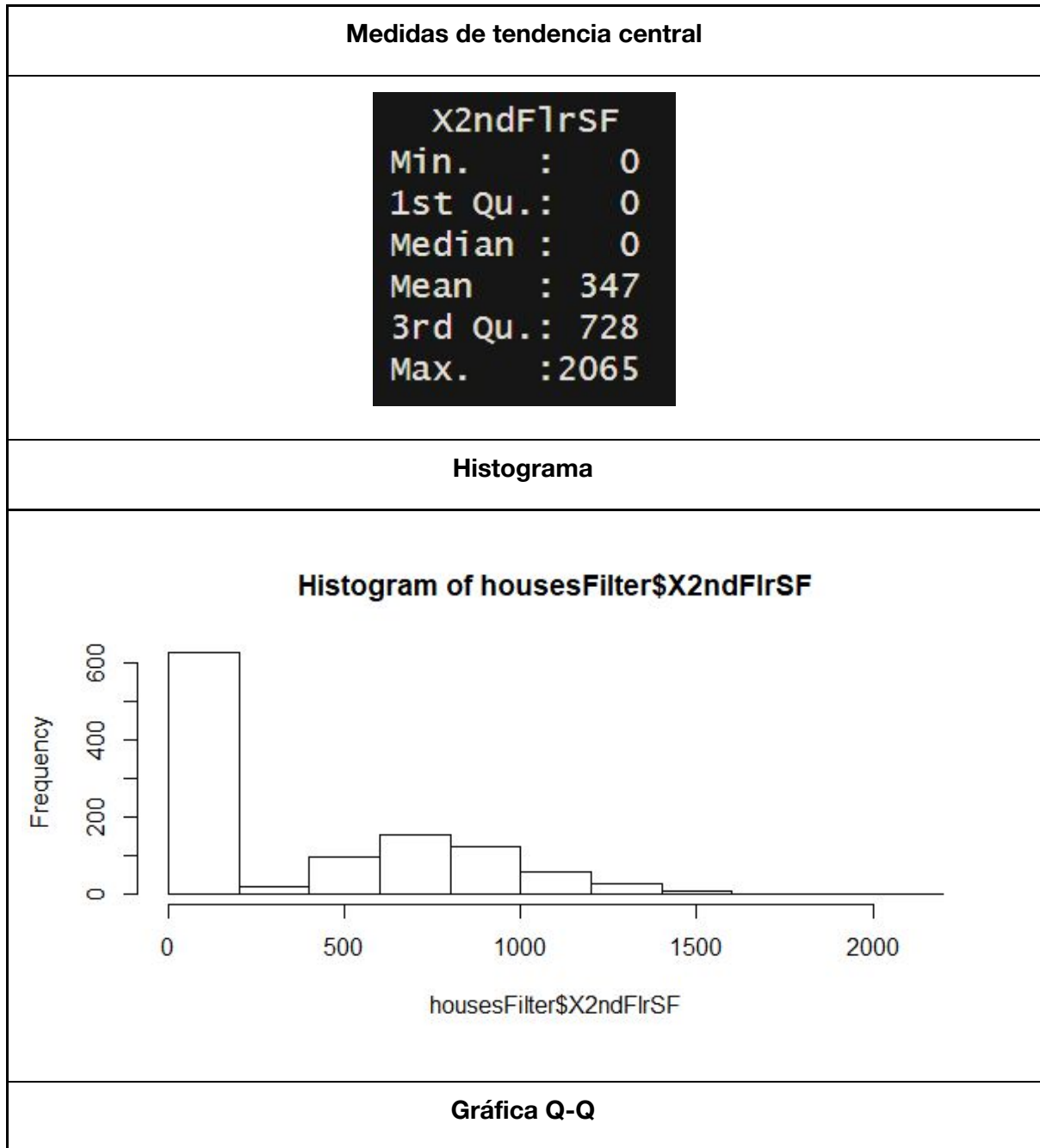


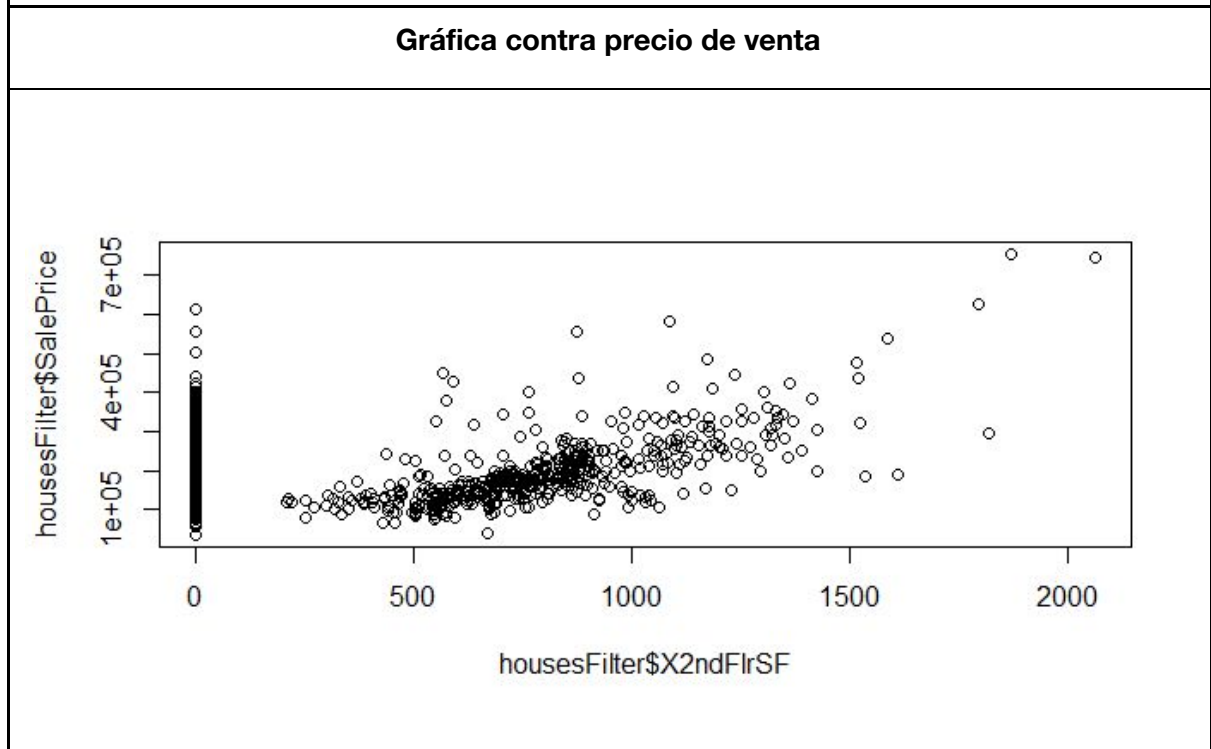
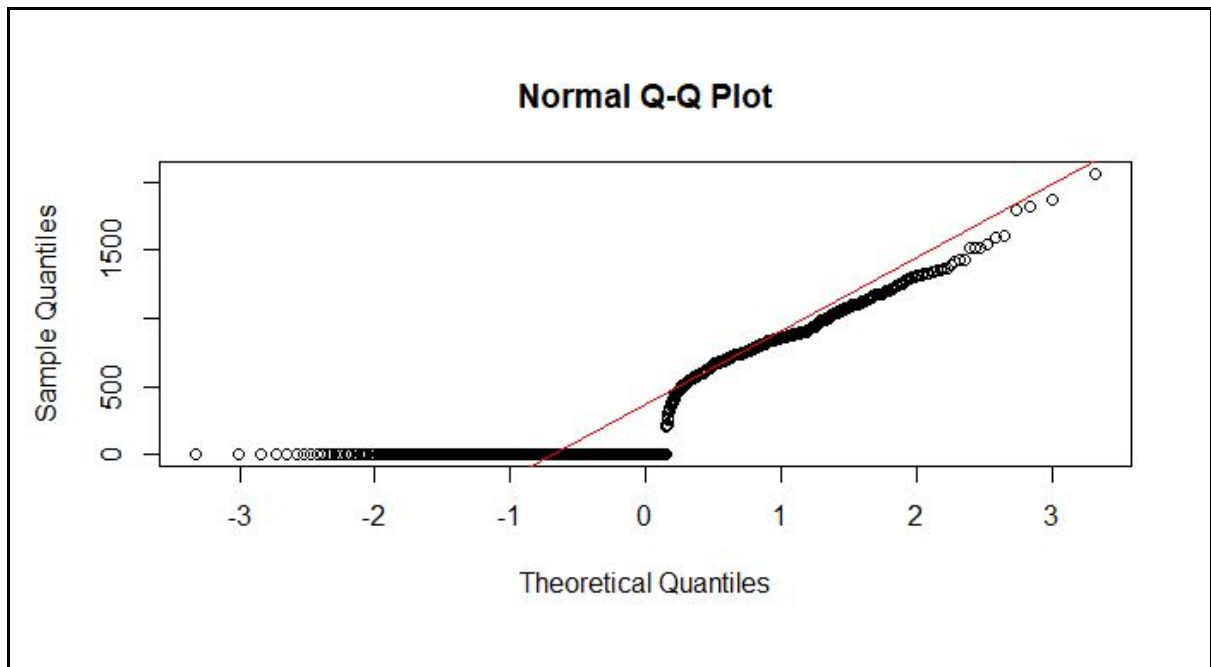
Gráfica contra precio de venta



- **X2ndFlrSF:** Metros cuadrados del segundo piso.

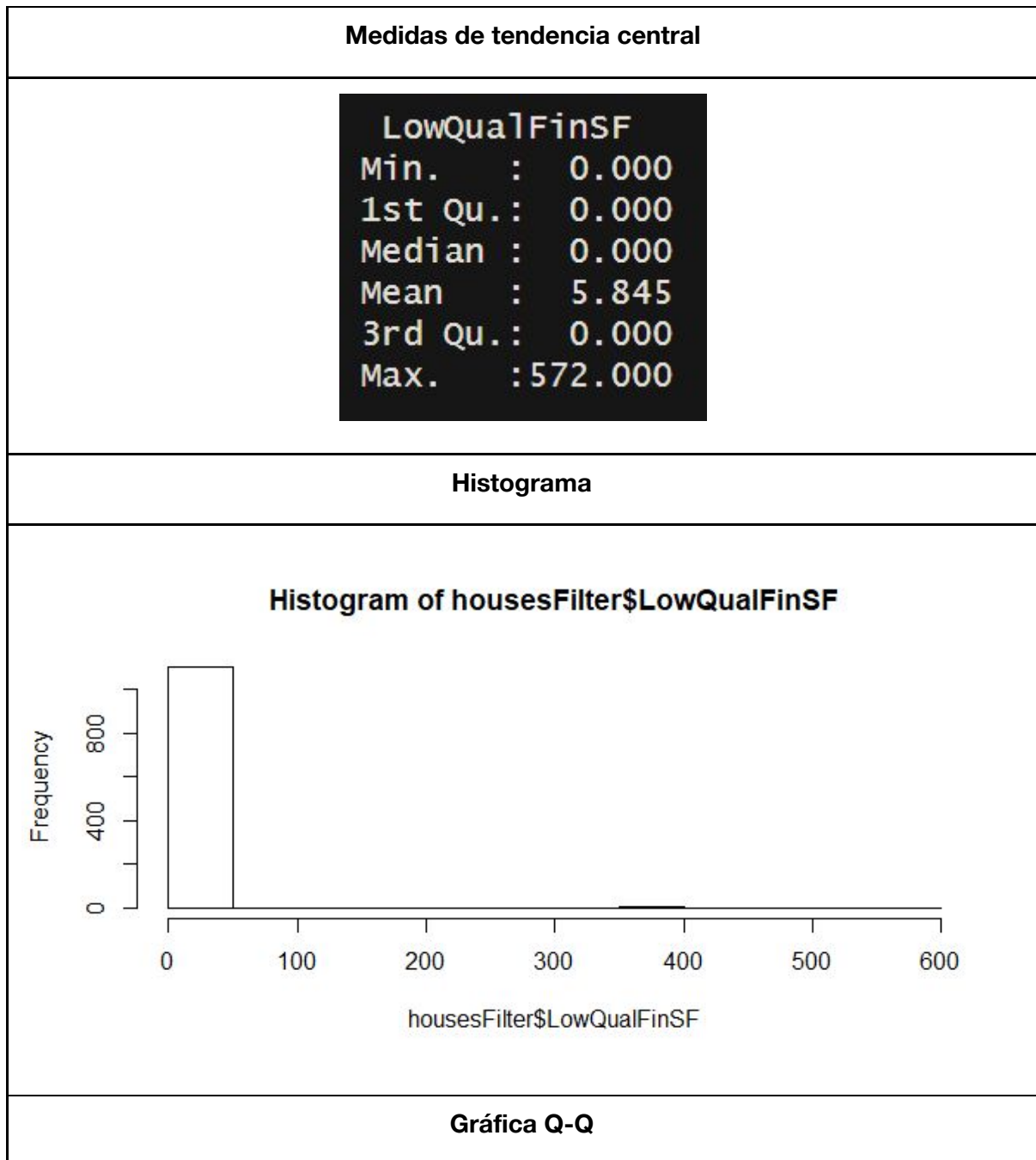
La variable X2dnFlrSF tiene simetría si se ignoran los valores 0's del dataset, los datos están ligeramente normalizados y tienen relación directa con el precio de venta.

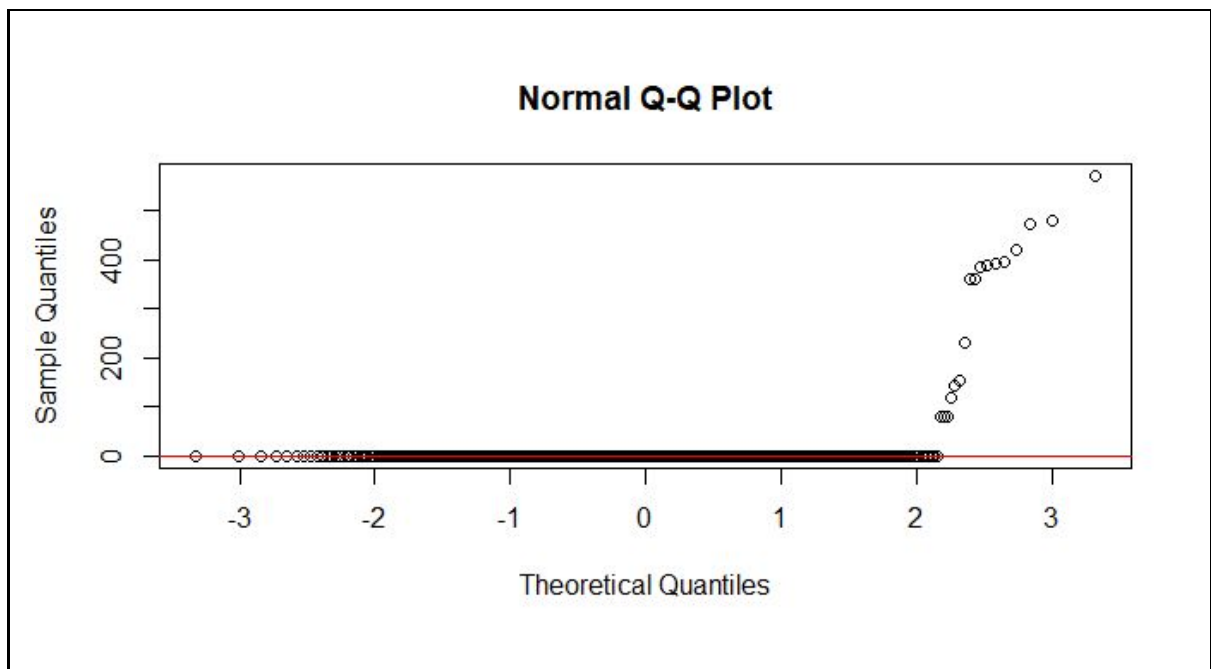




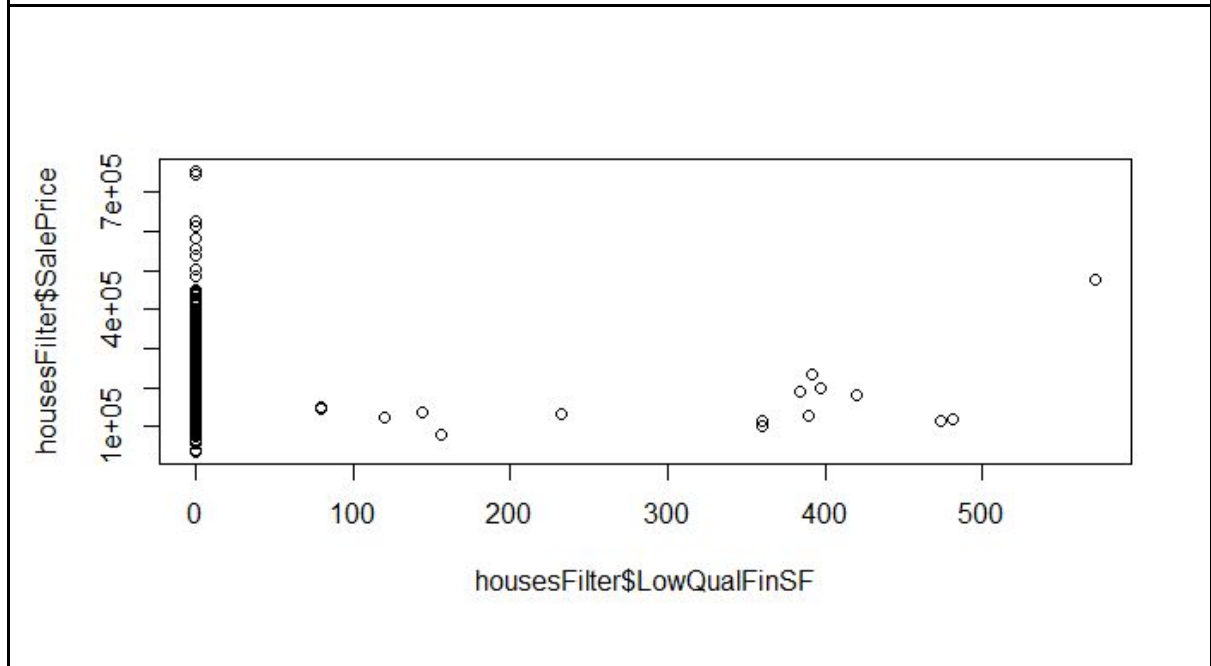
- **LowQualFinSF:** Cantidad de metros cuadrados de baja calidad.

Estos datos no muestran datos capaces de ser analizados.





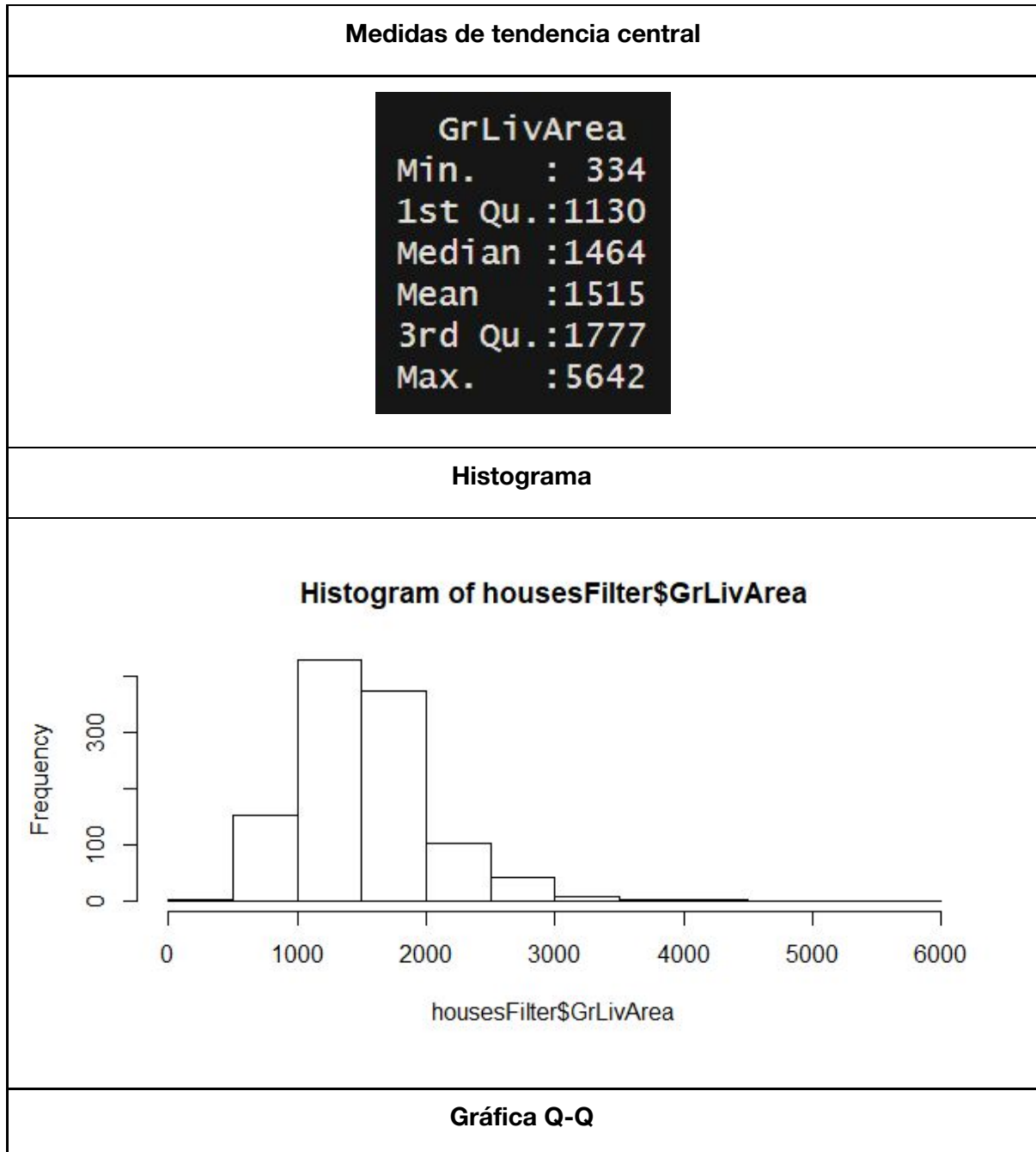
Gráfica contra precio de venta

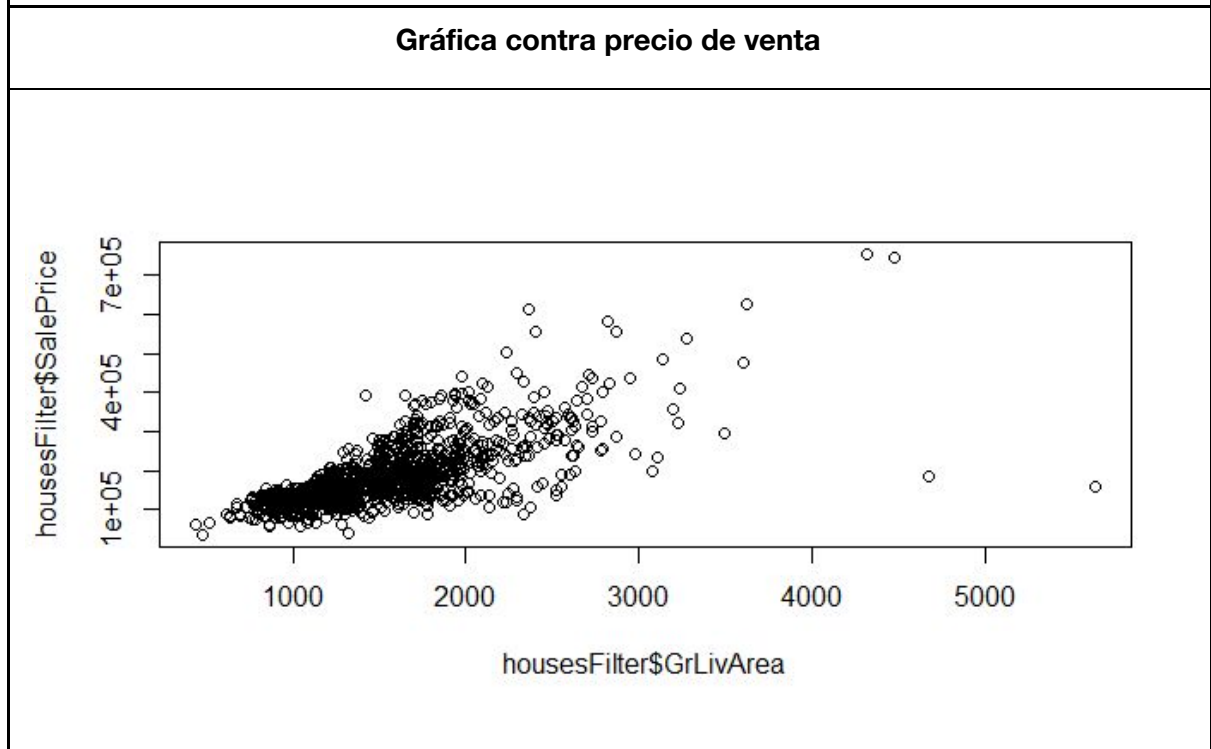
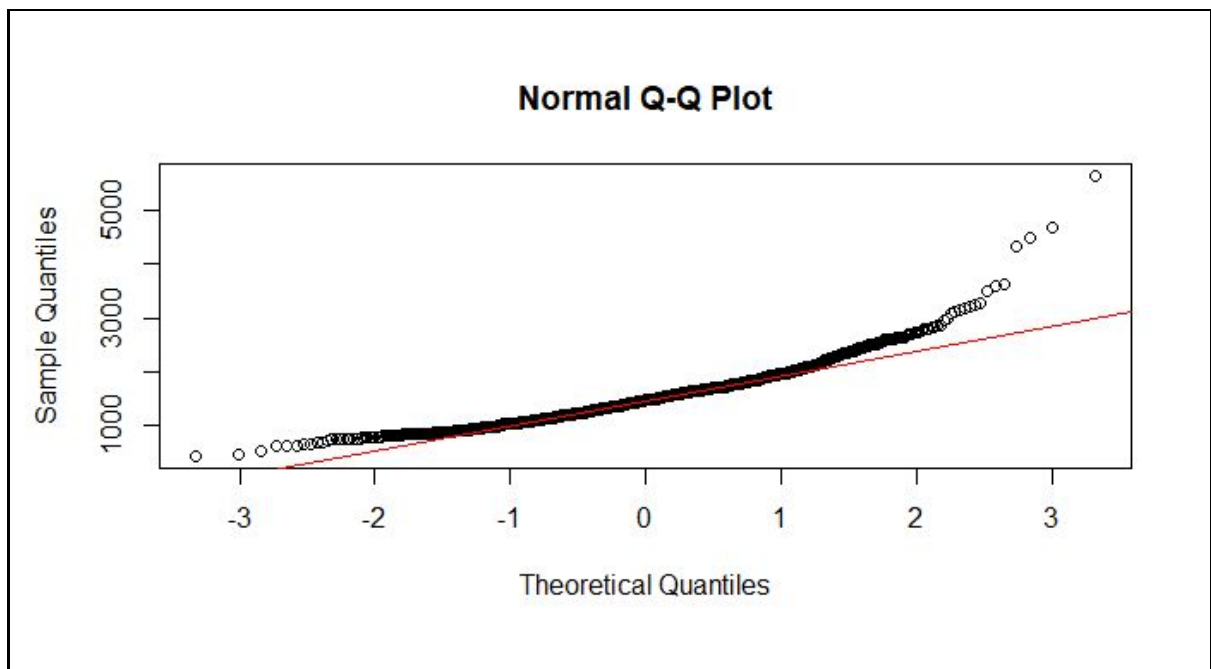




- **GrLivArea:** Área del primer piso.

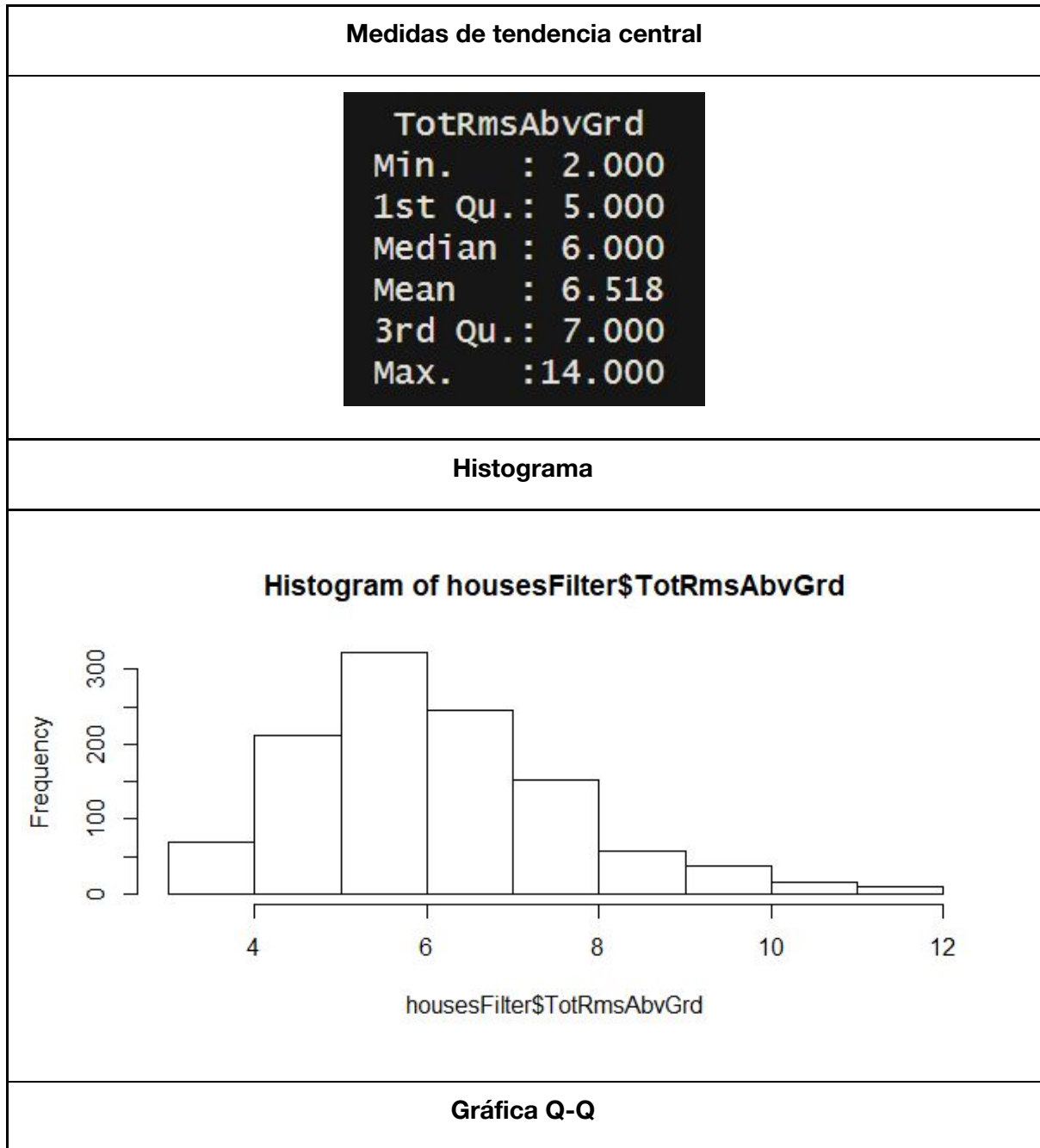
La variable GrLivArea tiene simetría, los datos están normalizados y tienen relación directa con el precio de venta.

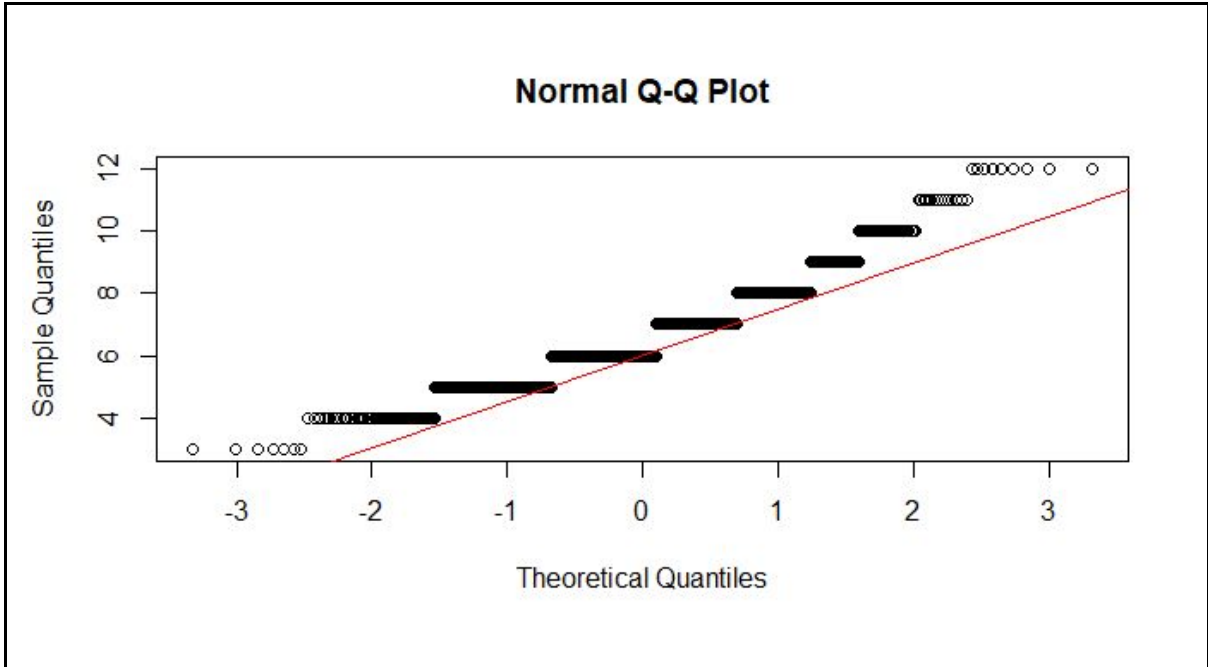




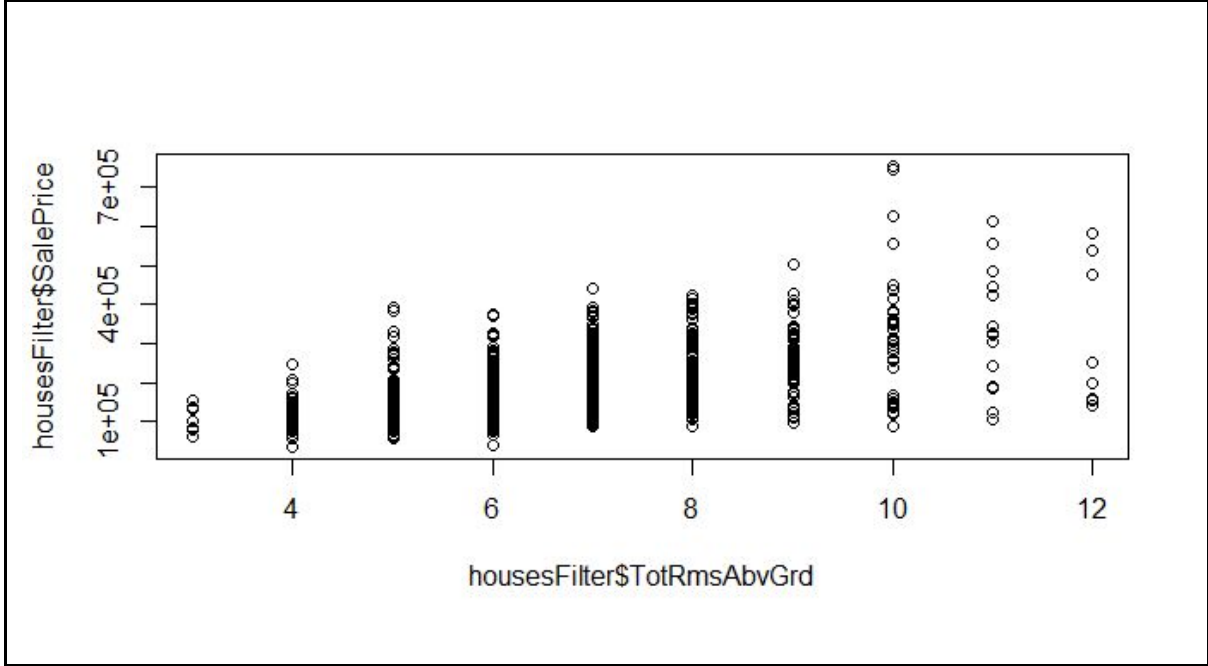
- **TotRmsAbvGrd:** Cantidad de metros cuadrados de área del sótano.

La variable TotRmsAbvGrd tiene simetría, sin embargo por ser datos categóricos no se puede analizar más allá de esto.



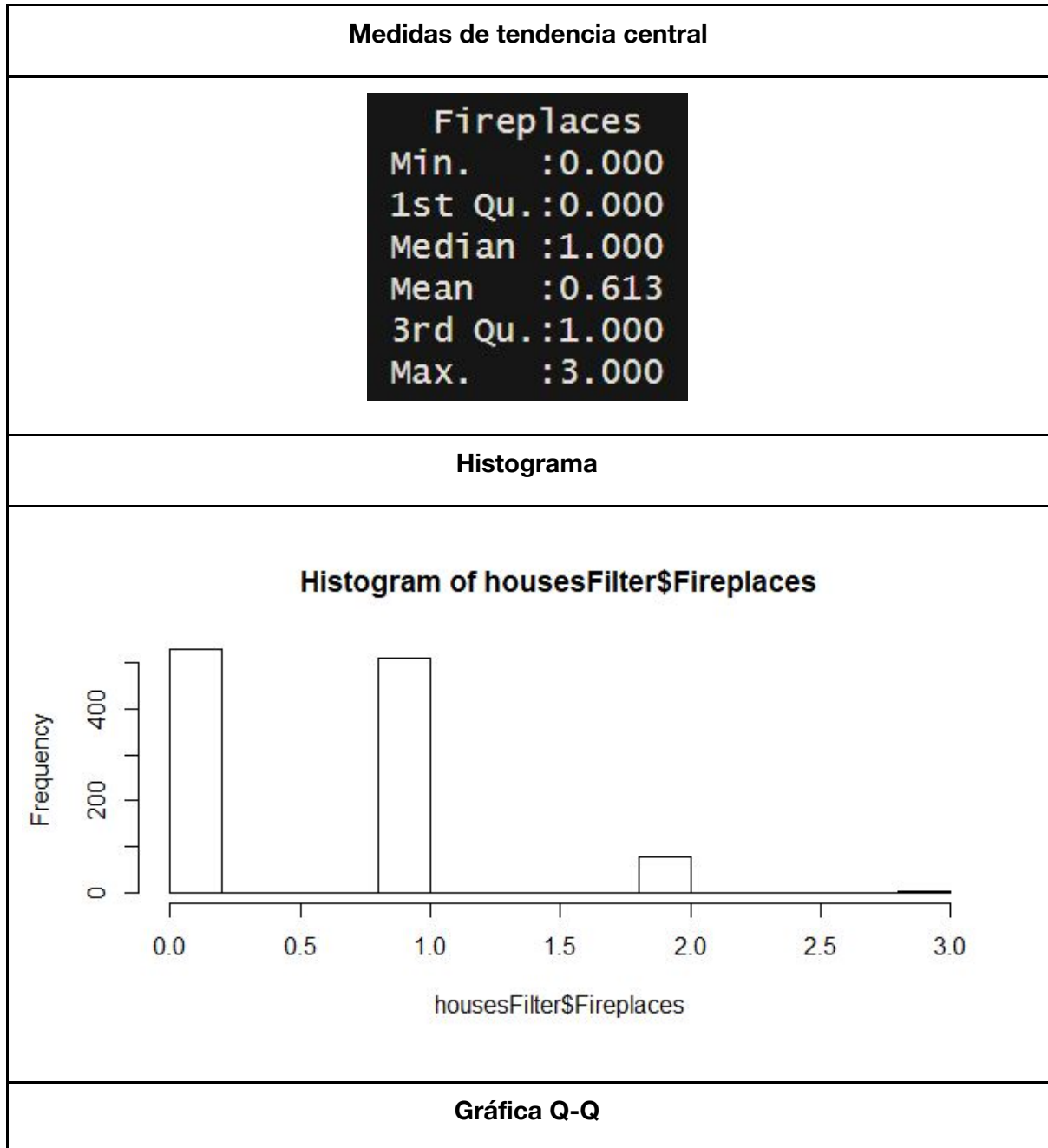


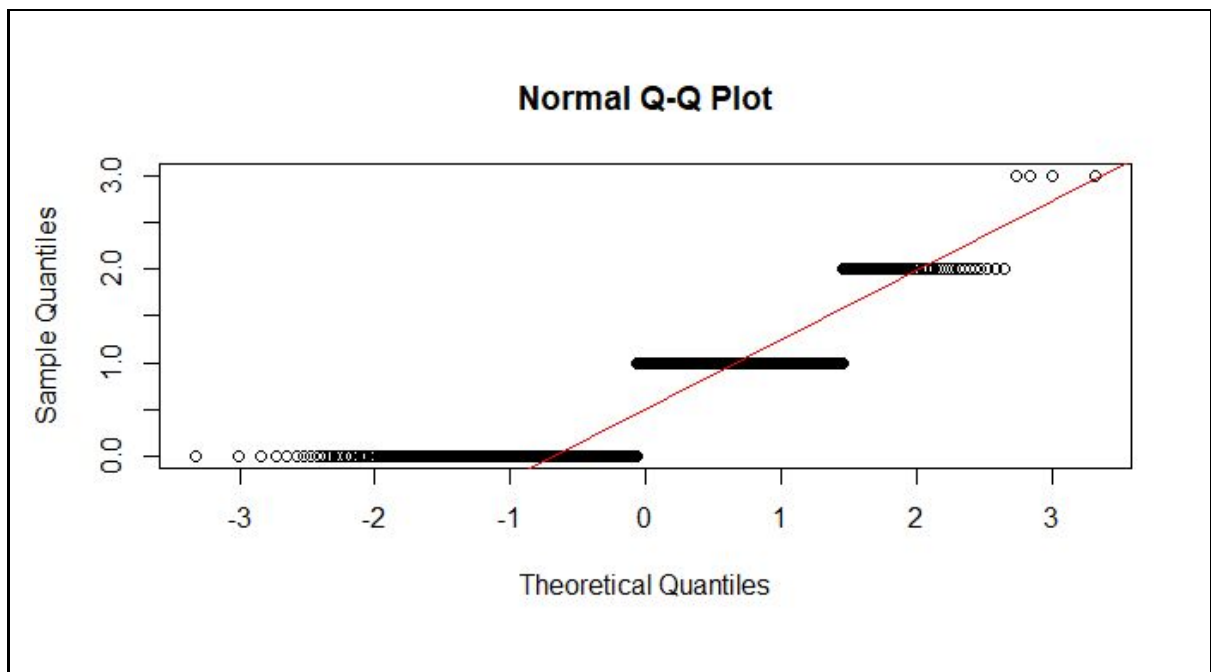
Gráfica contra precio de venta



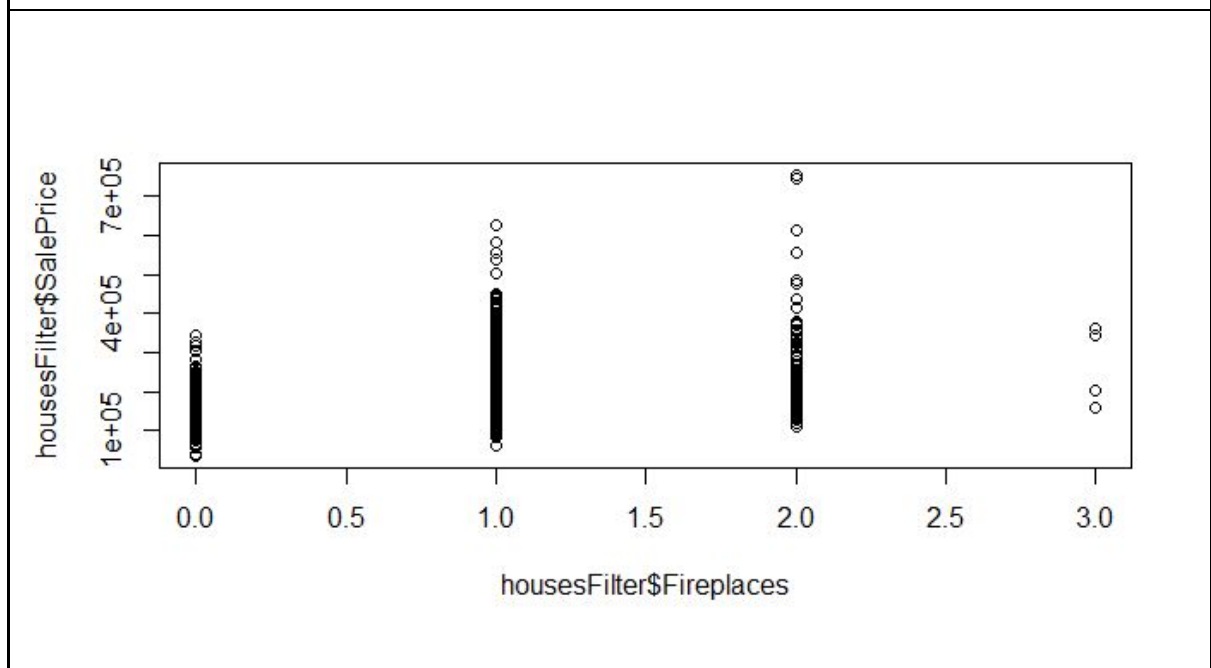
- **Fireplaces:** Número de chimeneas.

La variable Fireplaces no tiene simetría, y por ser datos categóricos no se puede analizar más allá de esto





Gráfica contra precio de venta



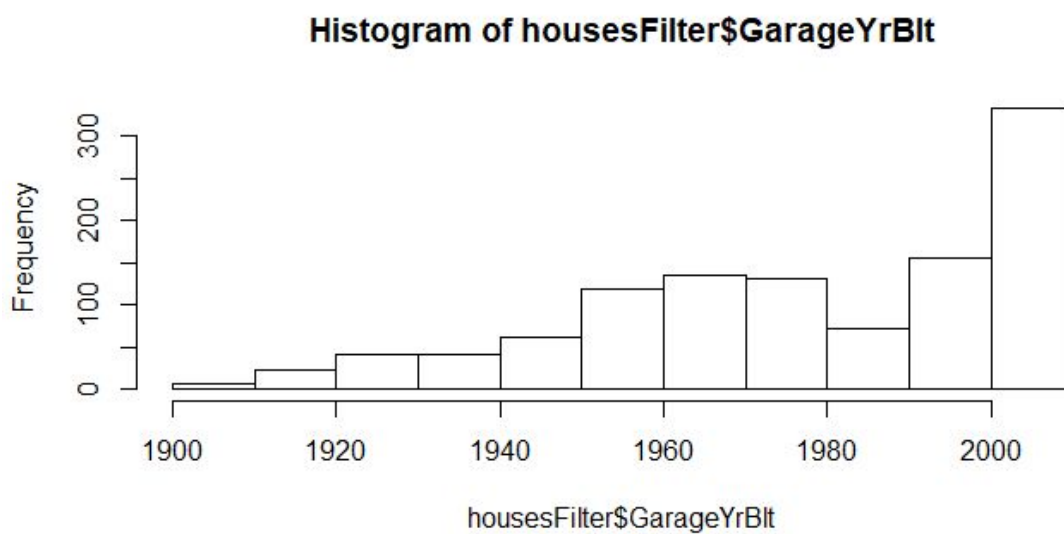
- **GarageYrBlt:** Año de construcción del garaje.

La variable GarageYrBlt tiene sesgo negativo, posee ligera normalización en sus datos mas no tiene relación con el precio.

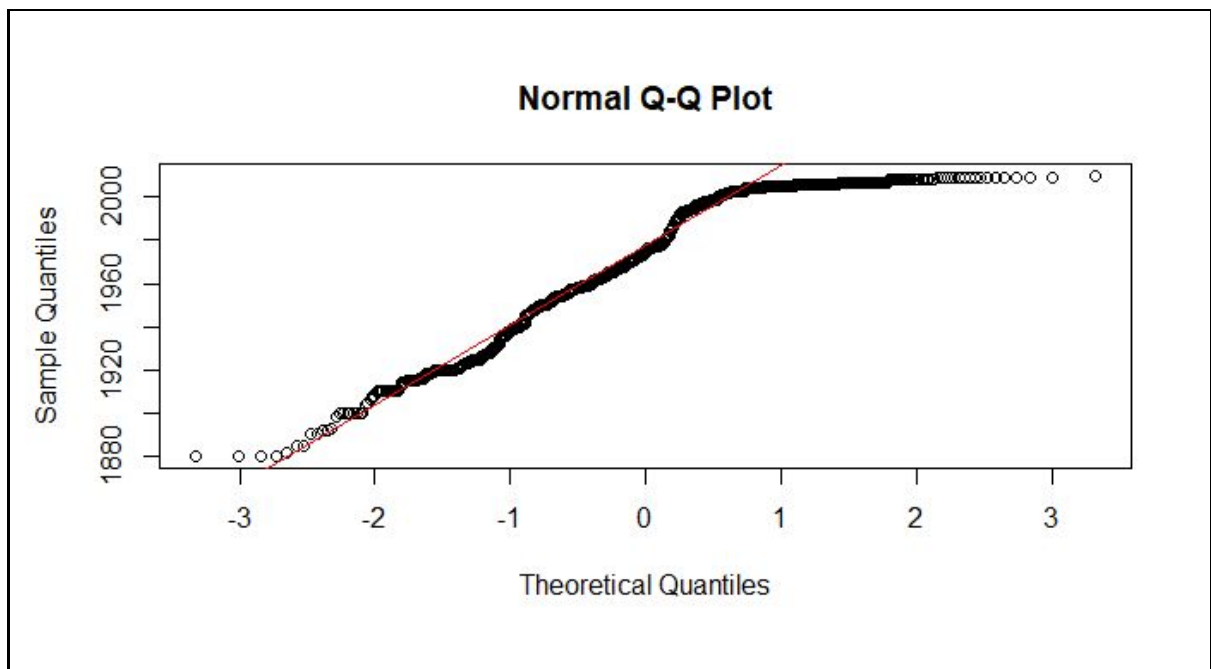
#### Medidas de tendencia central

```
GarageYrBlt
Min.      :1900
1st Qu.   :1961
Median    :1980
Mean      :1979
3rd Qu.   :2002
Max.      :2010
NA's      :81
```

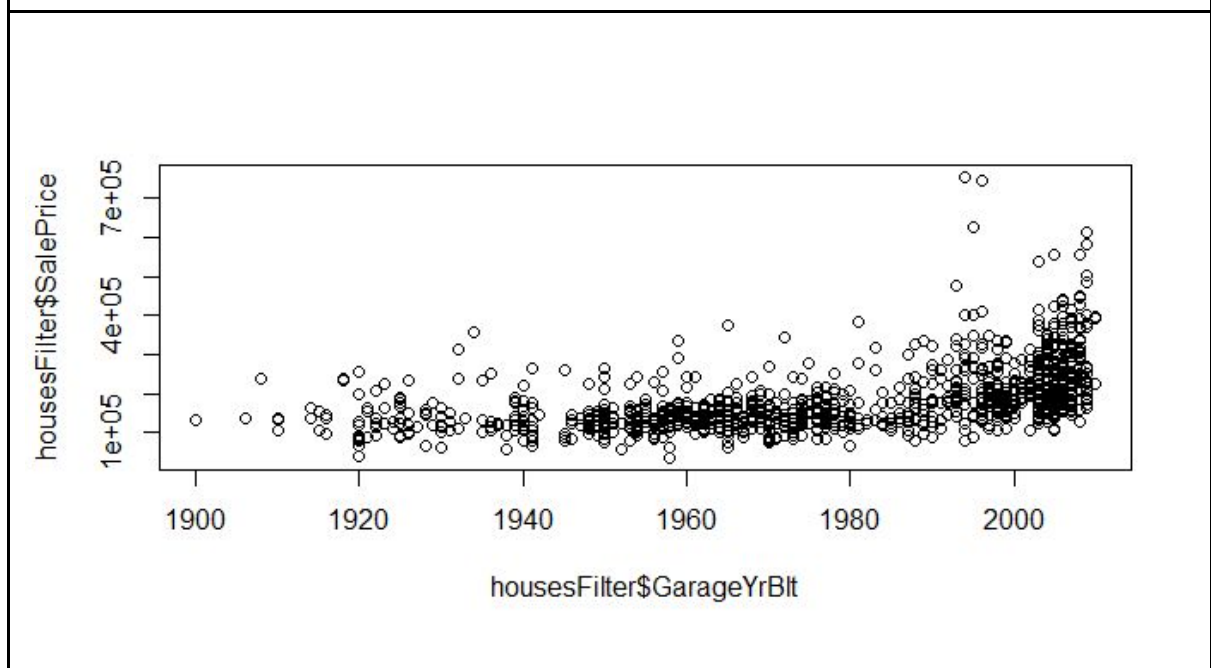
#### Histograma



#### Gráfica Q-Q



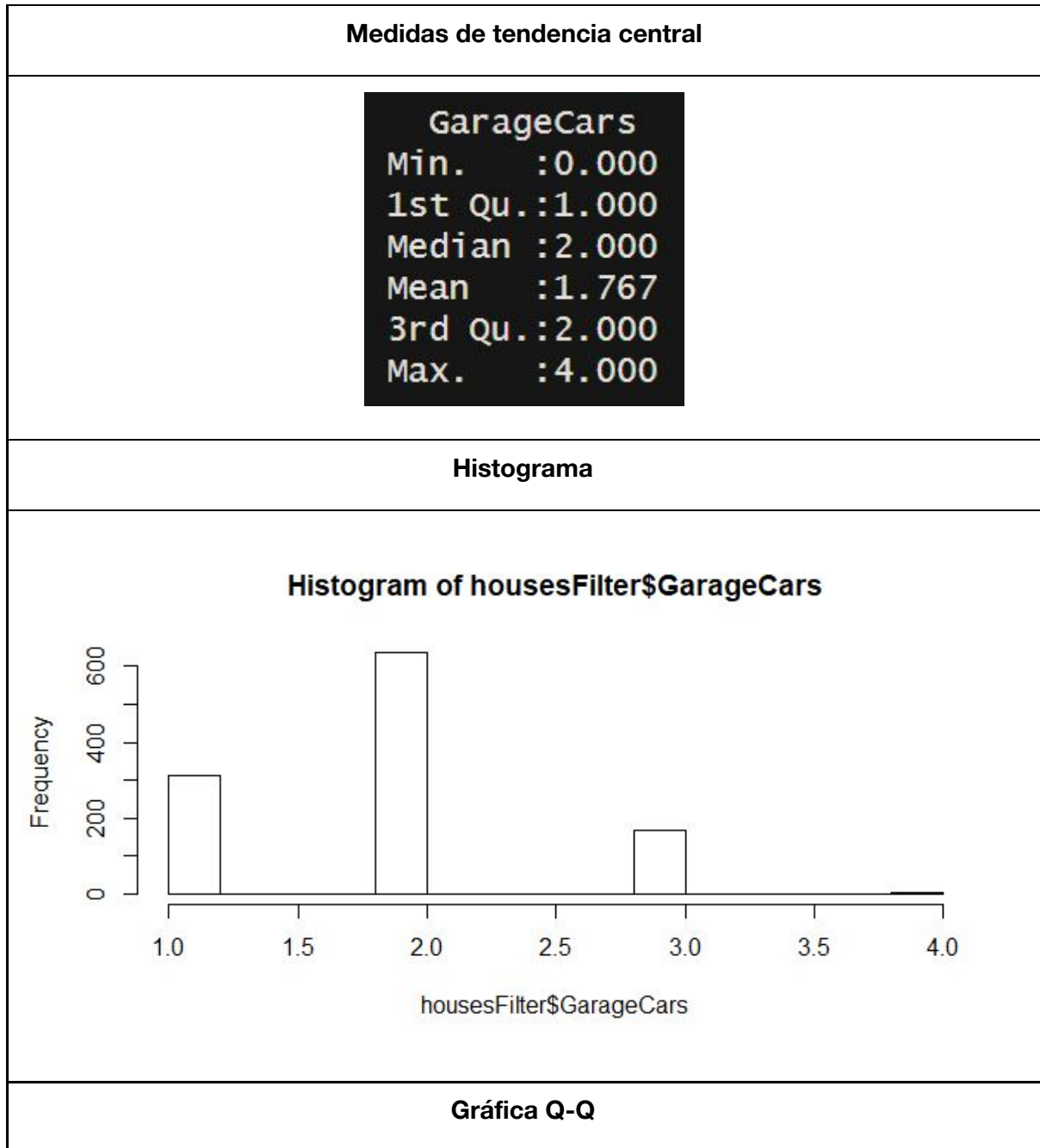
Gráfica contra precio de venta

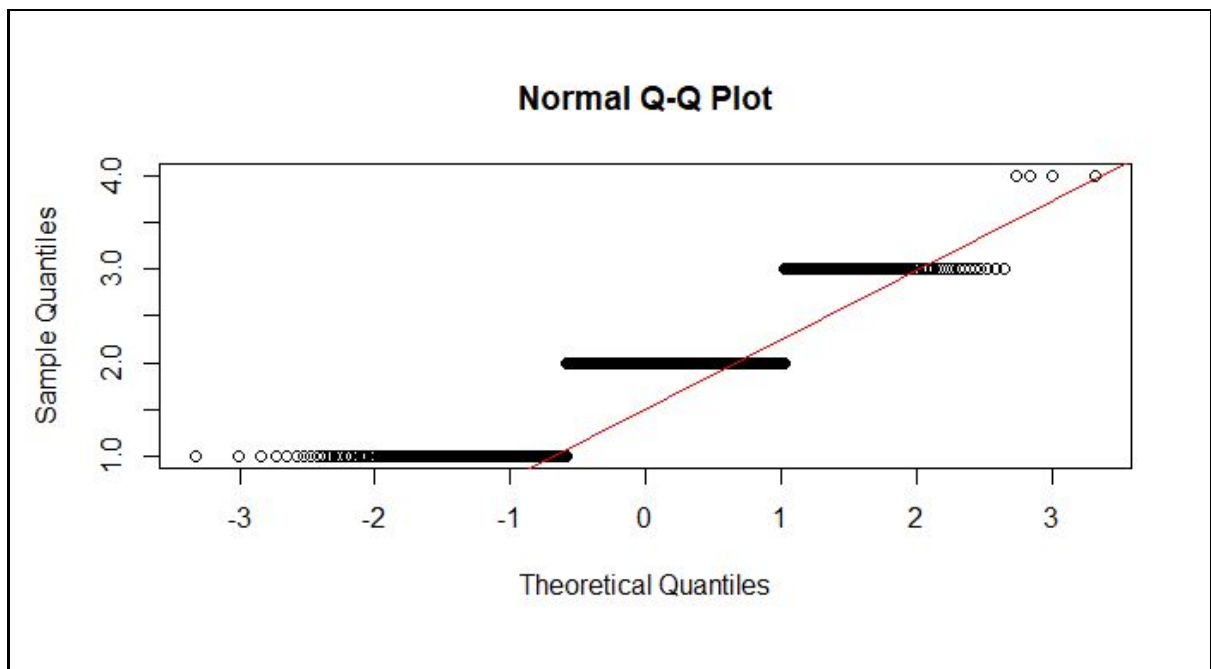




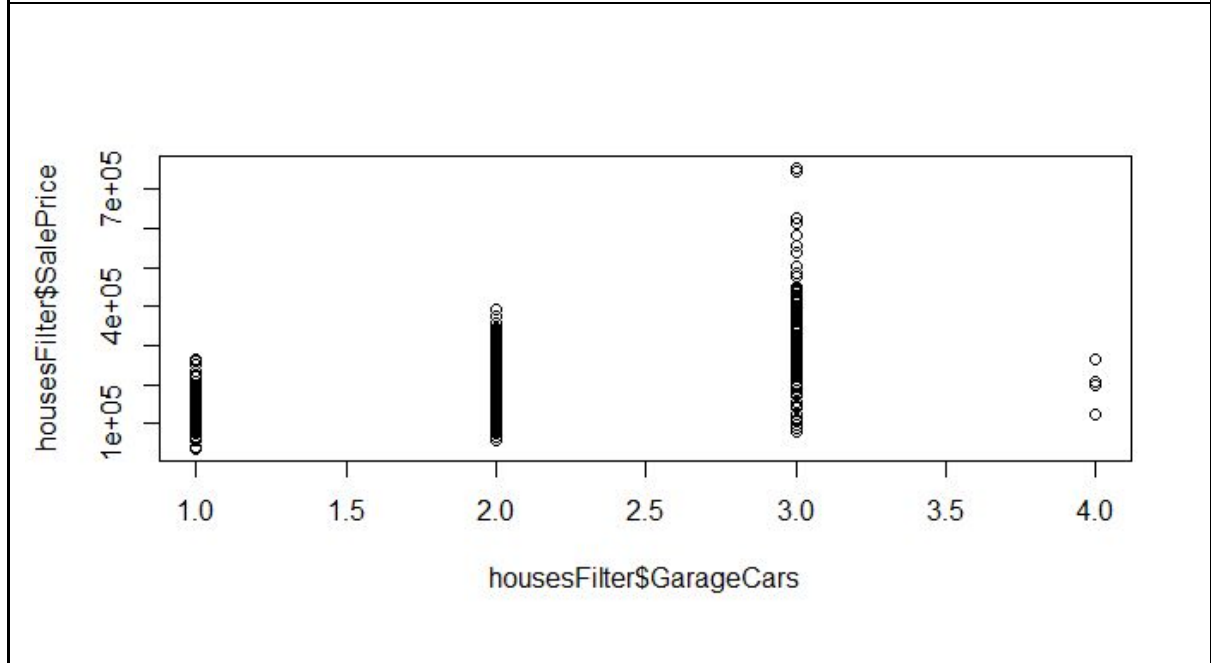
- **GarageCars:** Capacidad de carros del garaje.

La variable GarageCars no tiene simetría, y por ser datos categóricos no se puede analizar más allá de esto.





Gráfica contra precio de venta



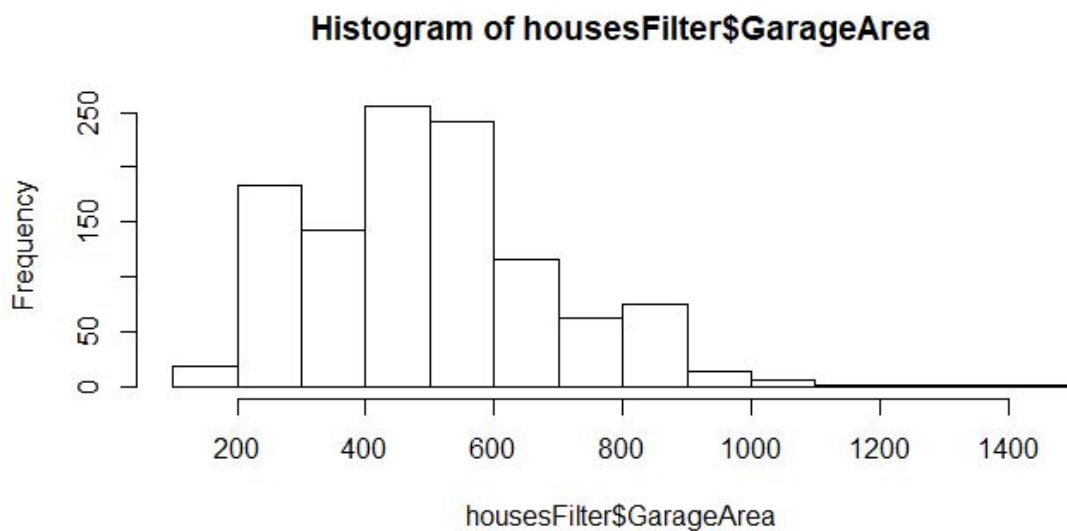
- **GarageArea:** Tamaño de garaje en metros cuadrados.

La variable GarageArea tiene ligera simetría, sus datos sí están normalizados y sí tienen relación con el precio.

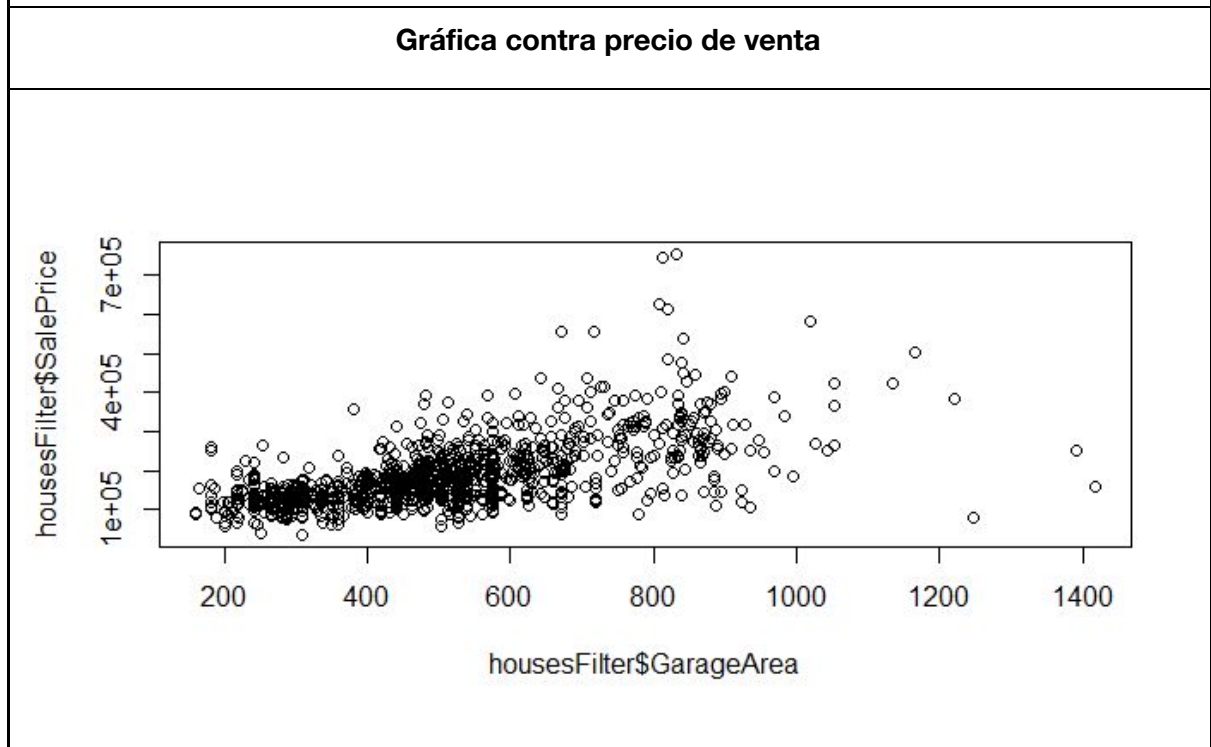
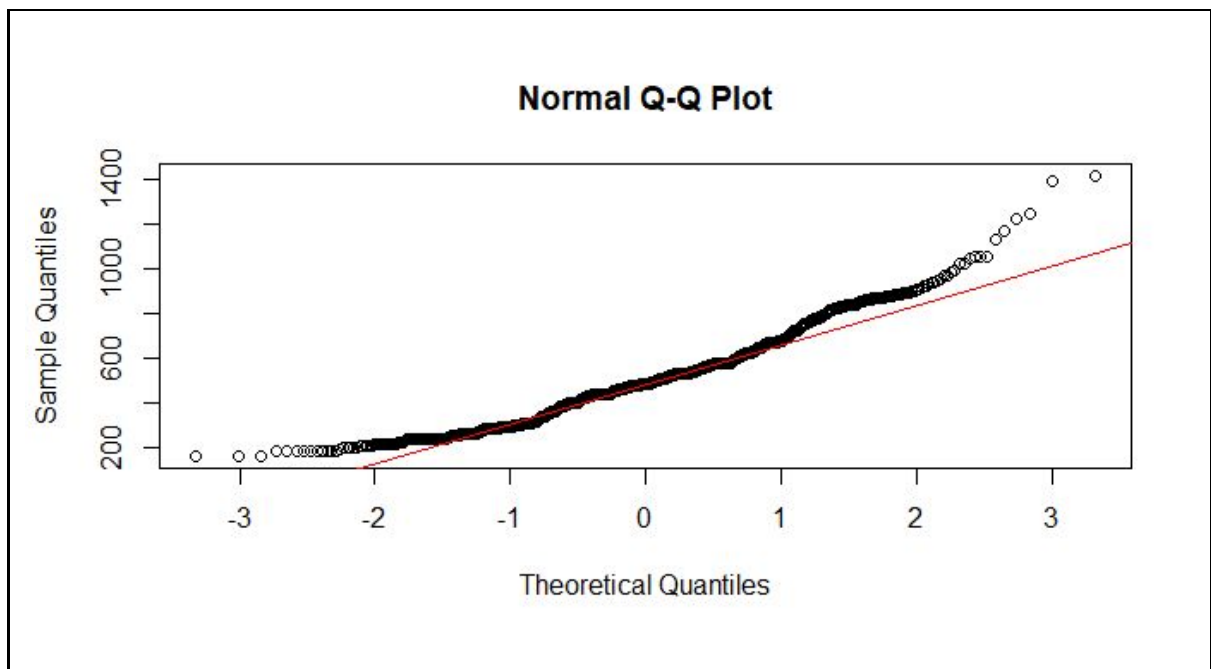
#### Medidas de tendencia central

```
GarageArea
Min.      :  0.0
1st Qu.   : 334.5
Median    : 480.0
Mean      : 473.0
3rd Qu.   : 576.0
Max.      :1418.0
```

#### Histograma

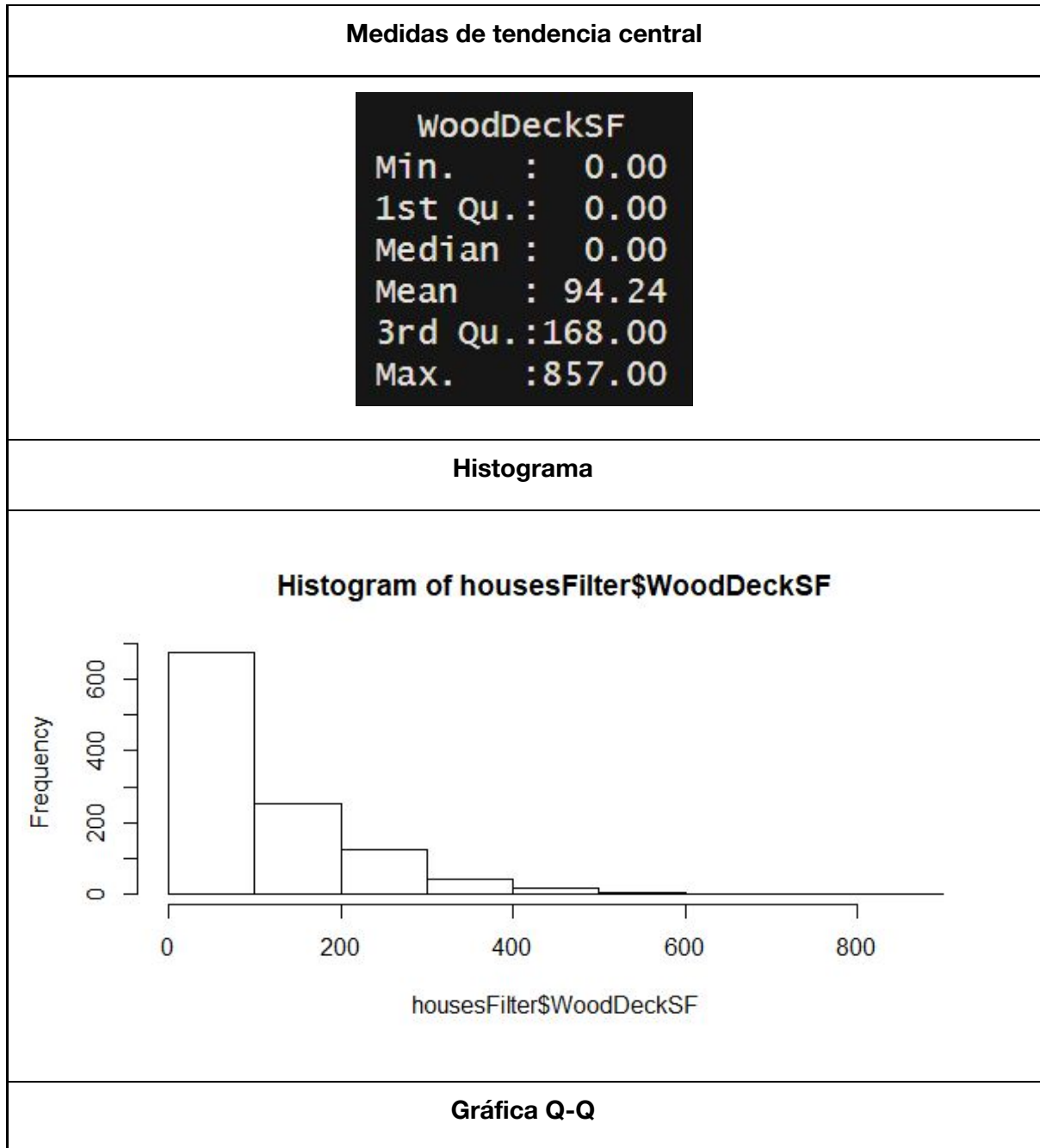


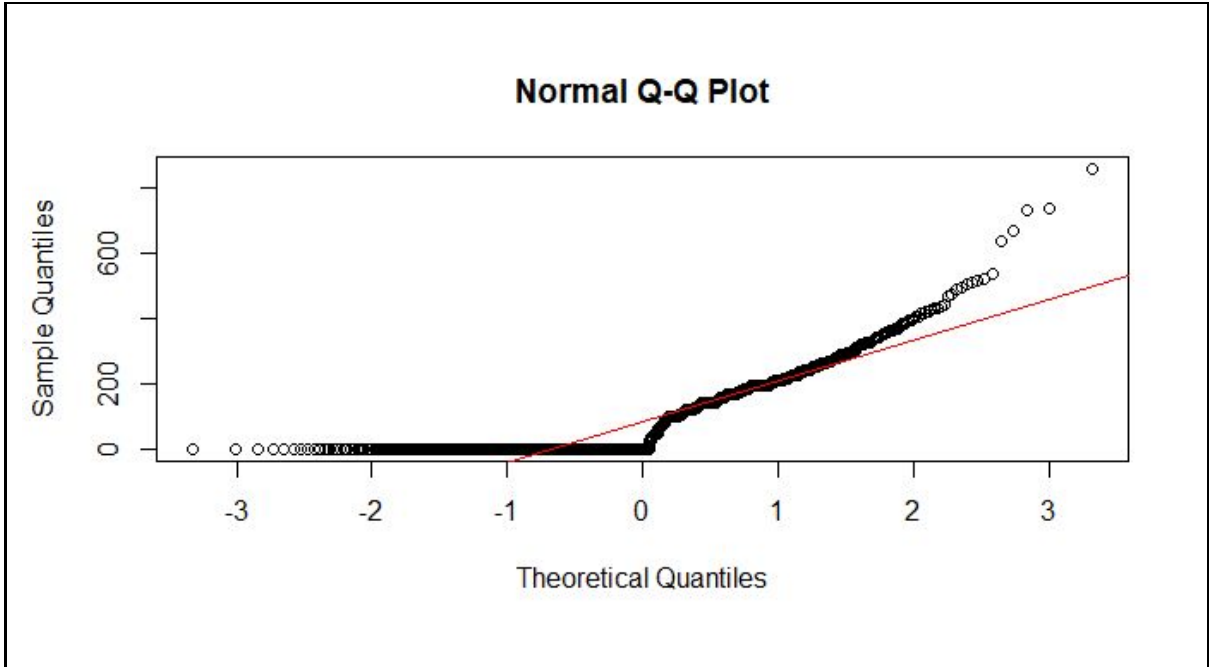
#### Gráfica Q-Q



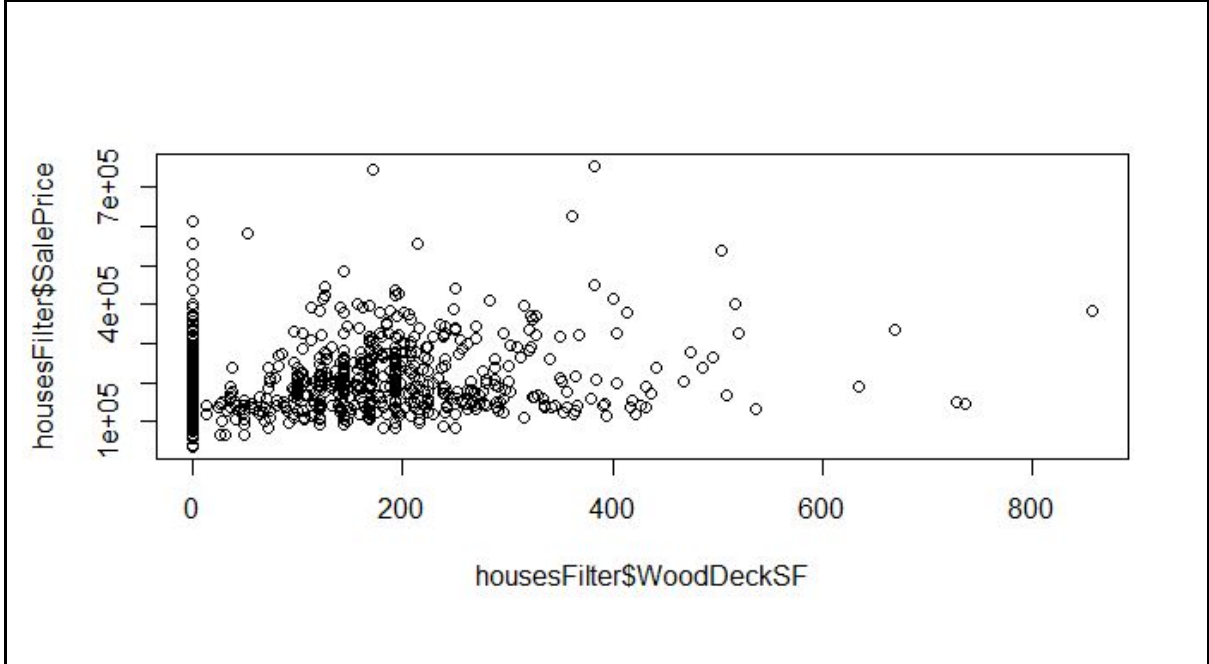
- **WoodDeckSF:** Área del deck de madera en metros cuadrados.

La variable WoodDeckSF tiene sesgo positivo, no tiene datos normalizados y tampoco relación con el precio de venta.



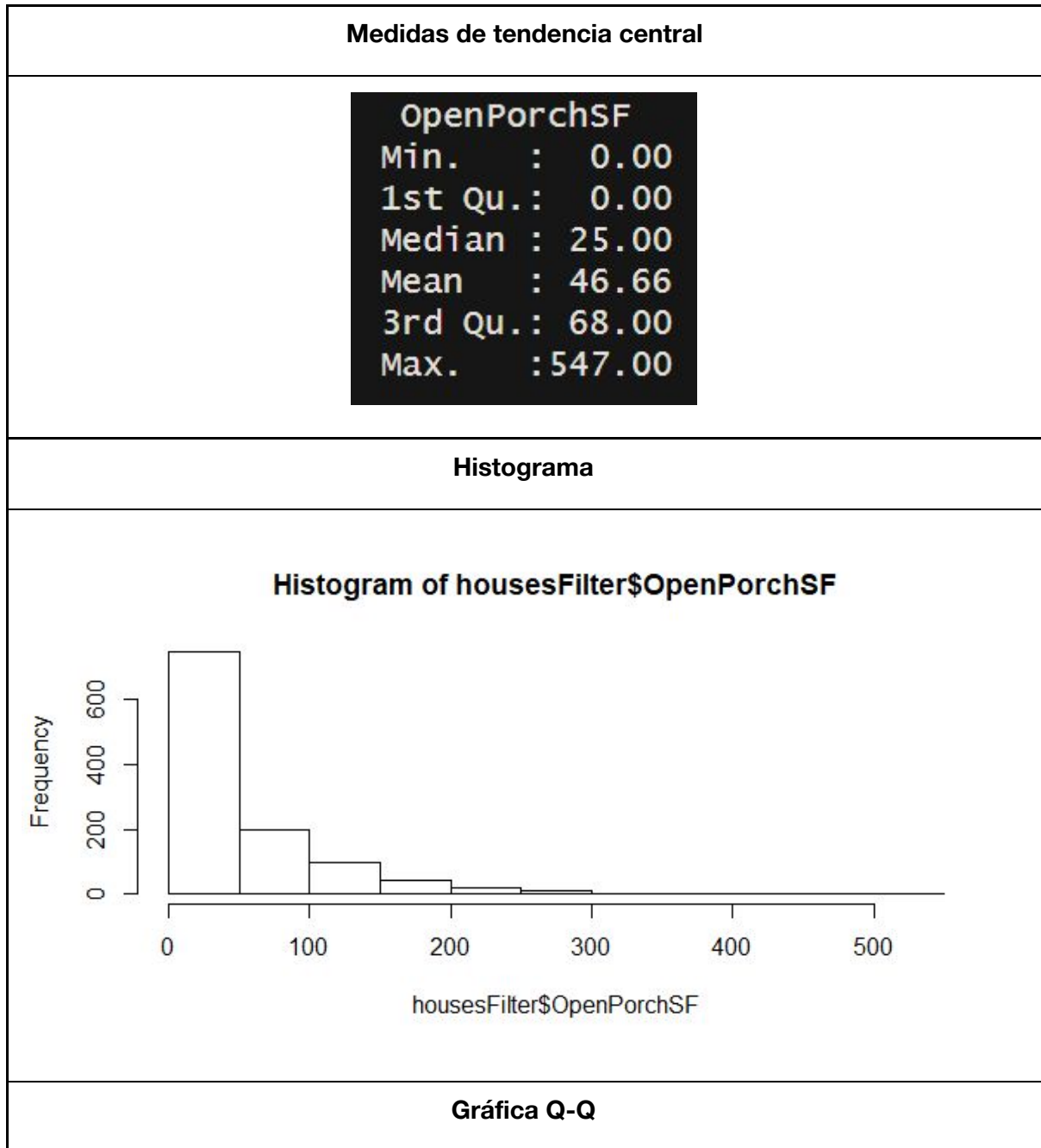


**Gráfica contra precio de venta**

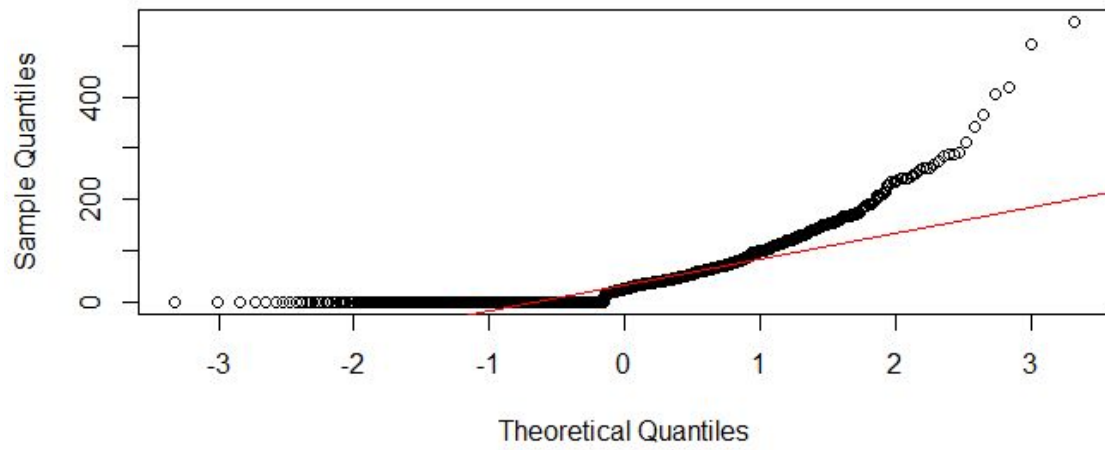


- **OpenPorchSF:** Área del porche abierta, en metros cuadrados.

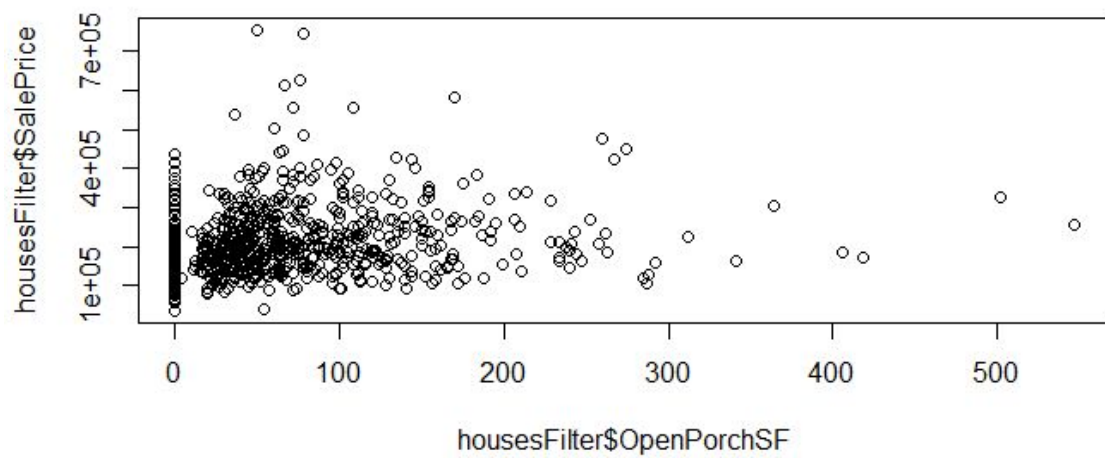
La variable TotRomsAbrGrd tiene sesgo positivo, no tiene simetría y tampoco relación con el precio de venta.



Normal Q-Q Plot



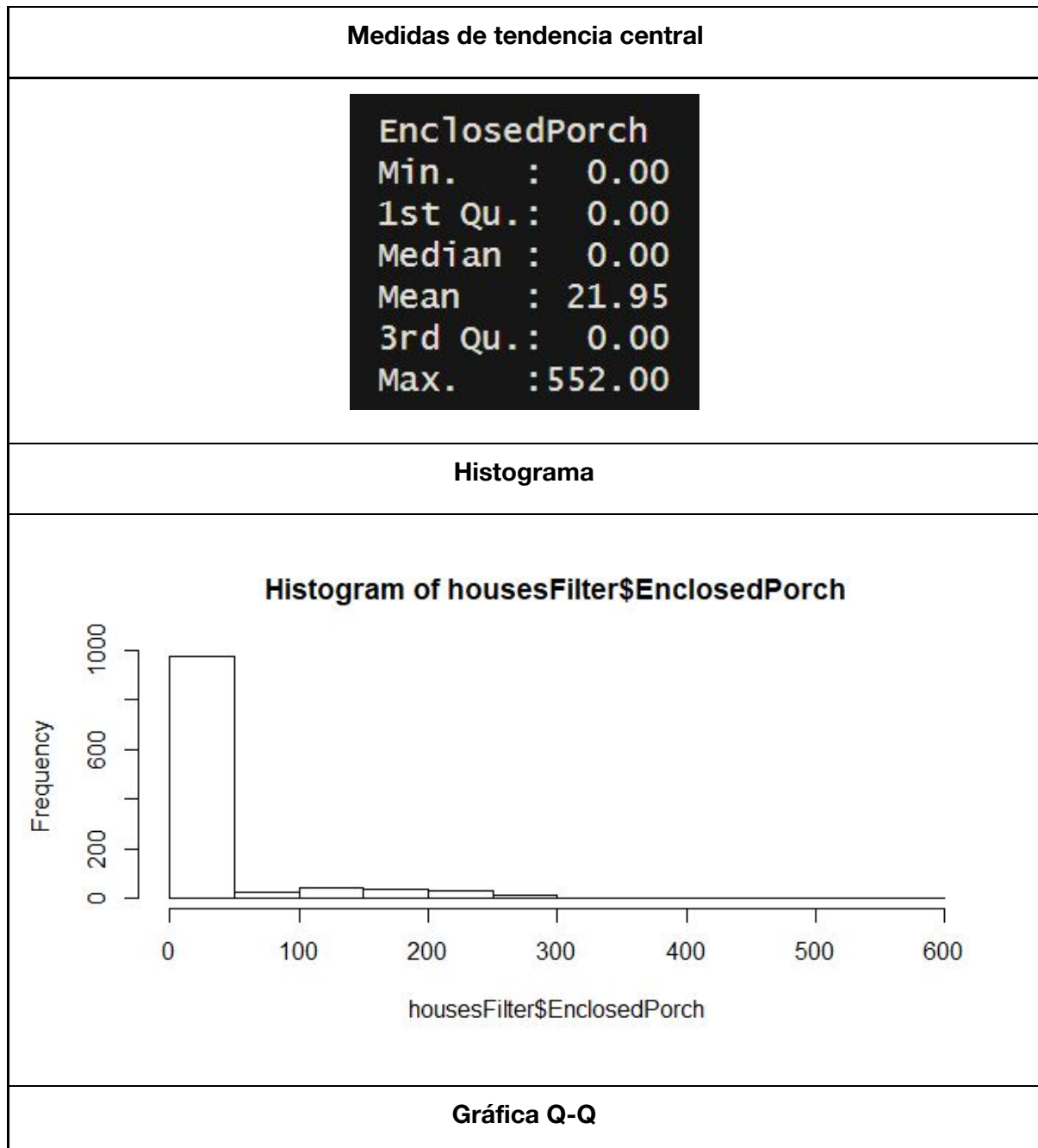
Gráfica contra precio de venta

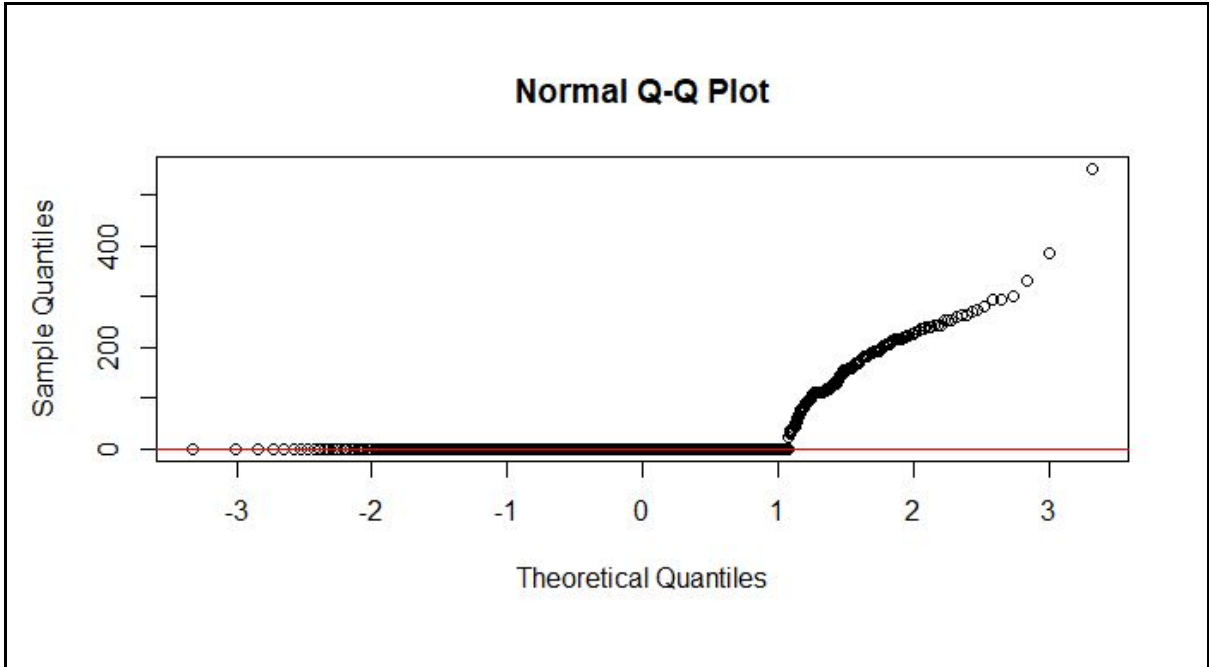




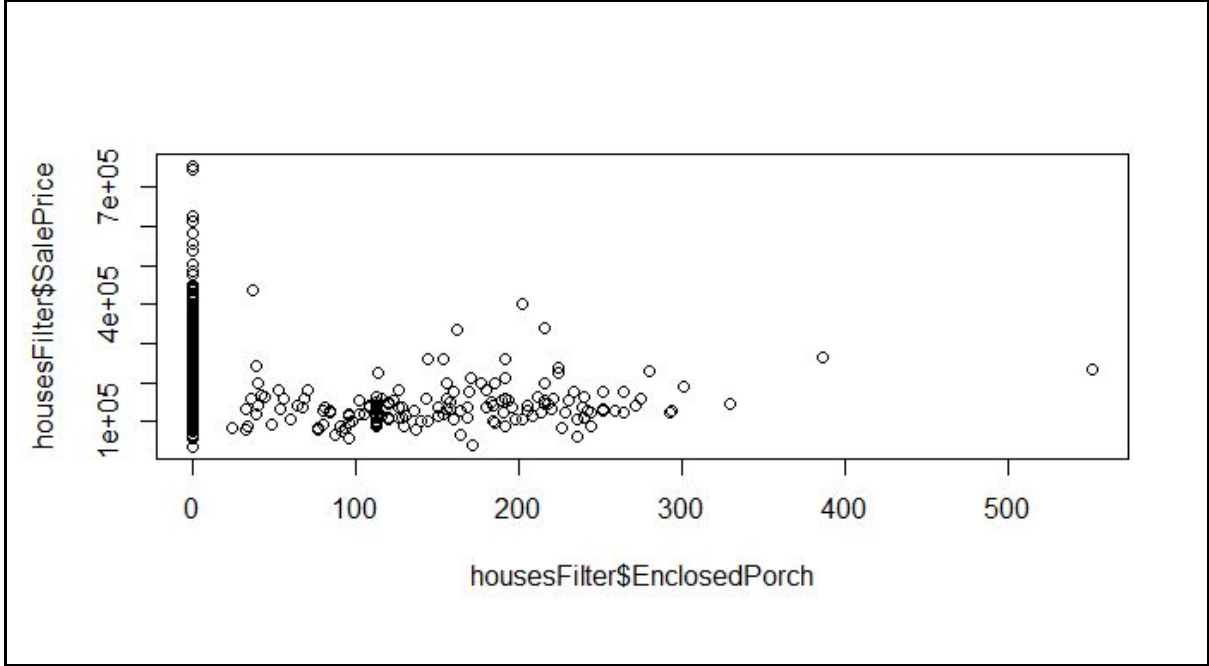
- **EnclosedPorch:** Área del porche cerrada, en metros cuadrados.

Estos datos no brindan información interesantes para analizar.



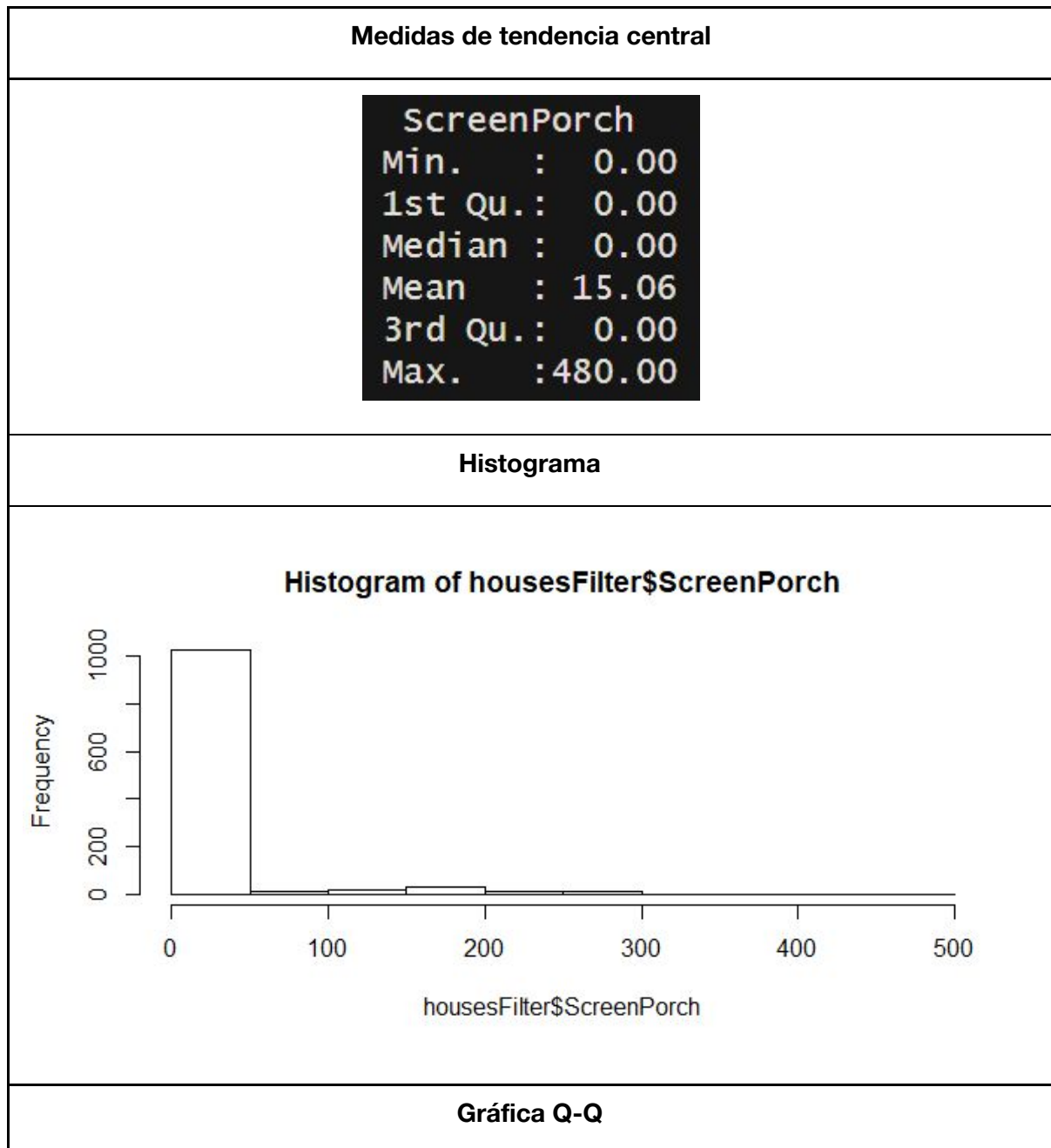


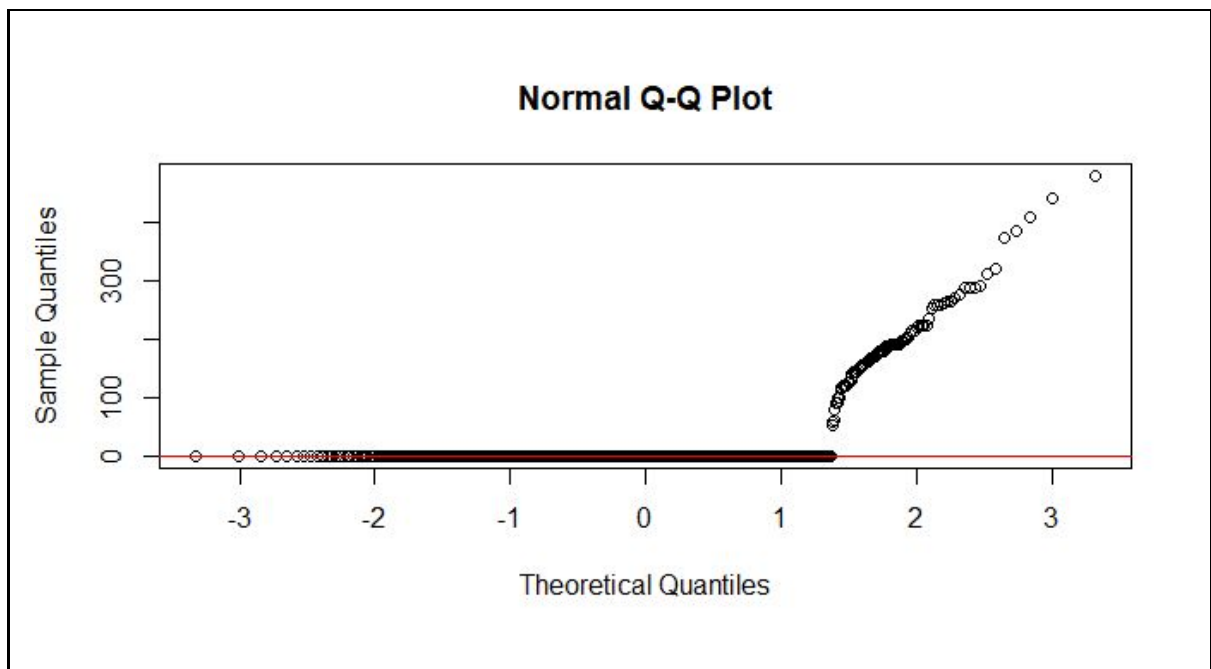
**Gráfica contra precio de venta**



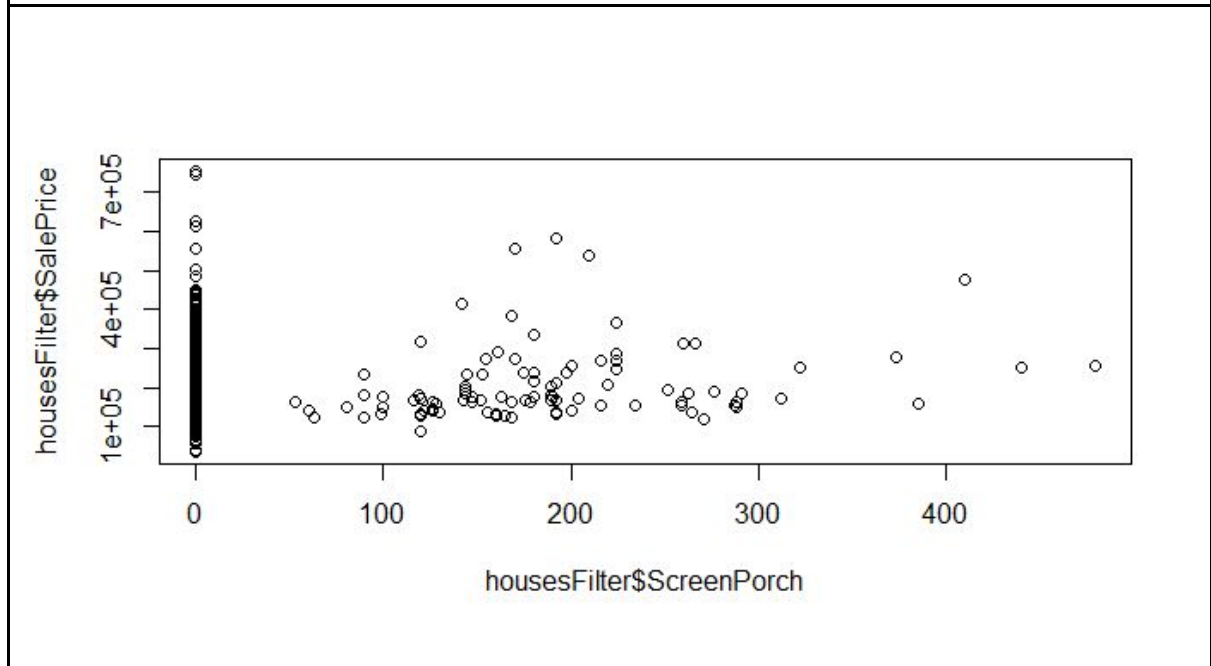
- **ScreenPorch:** Área del porche con pantalla, en metros cuadrados.

Estos datos no brindan información interesantes para analizar.



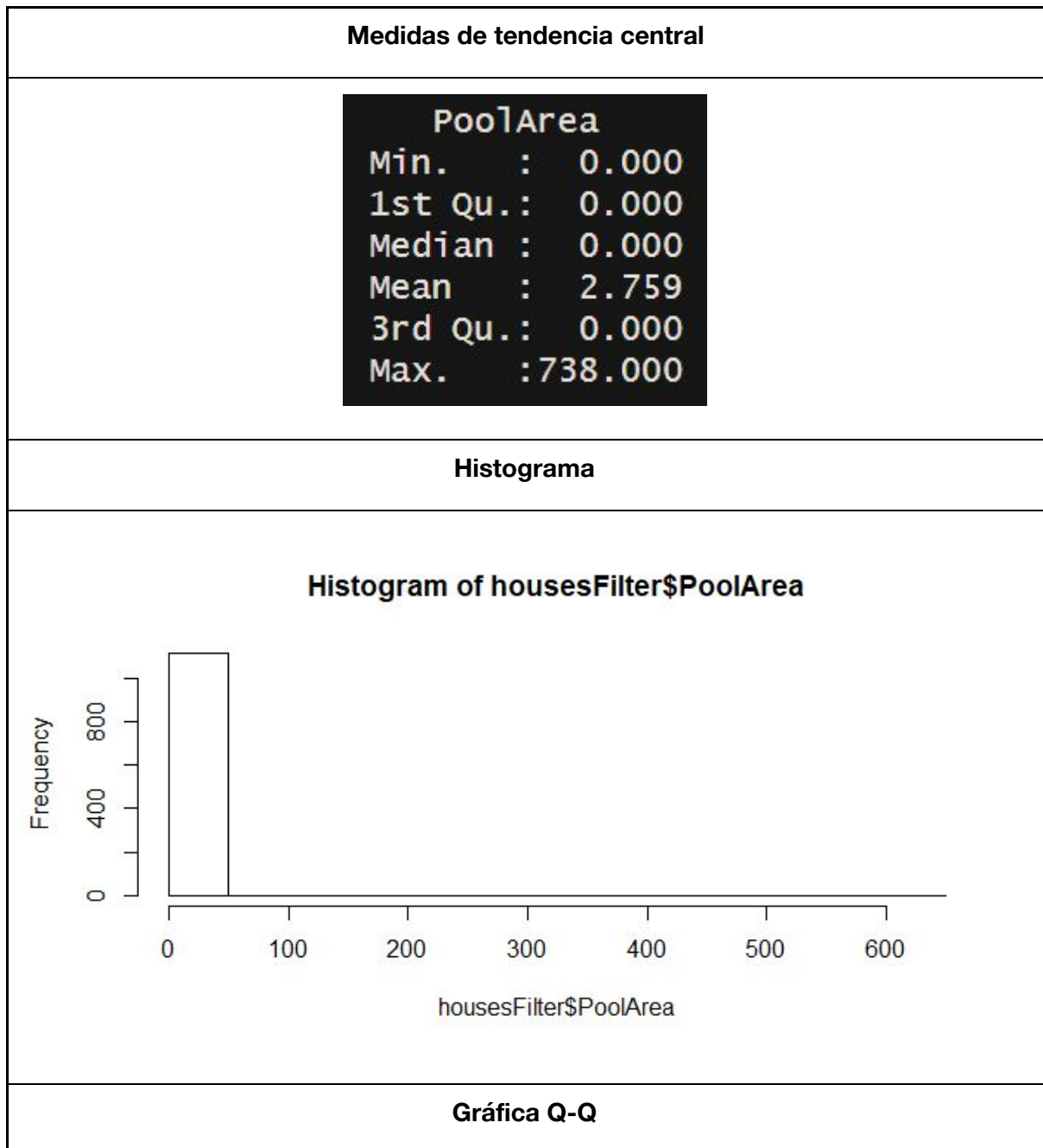


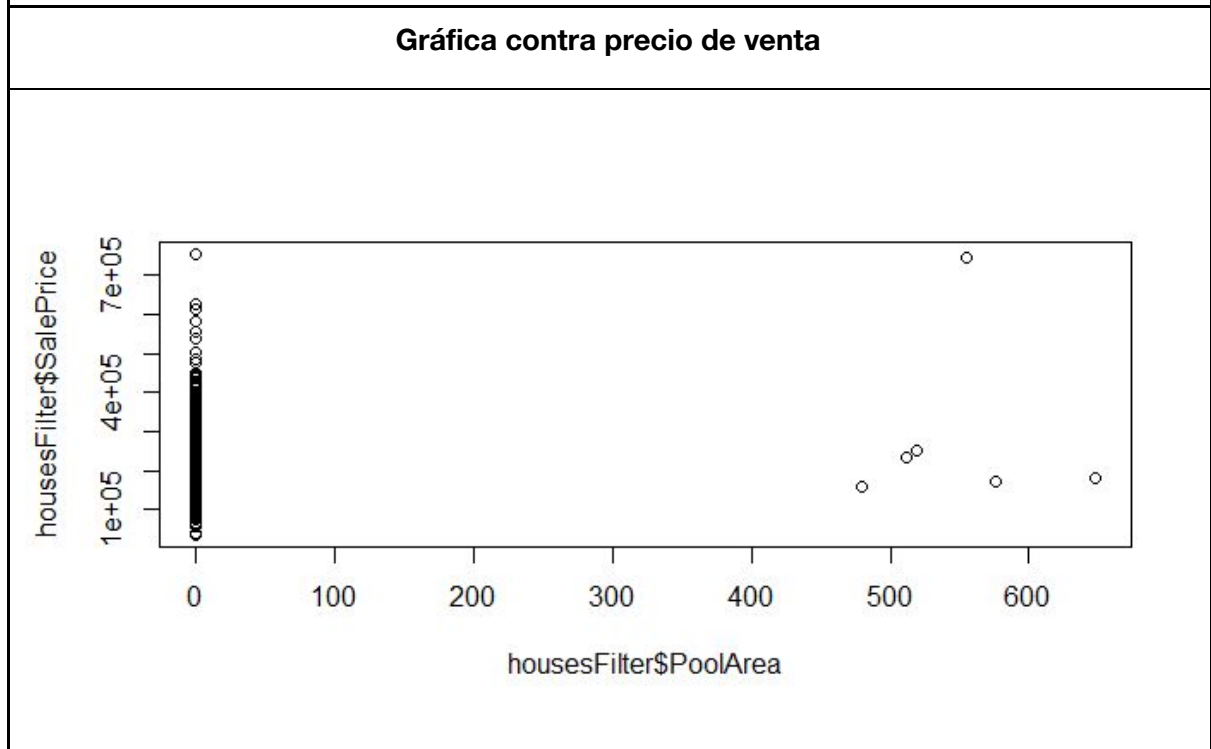
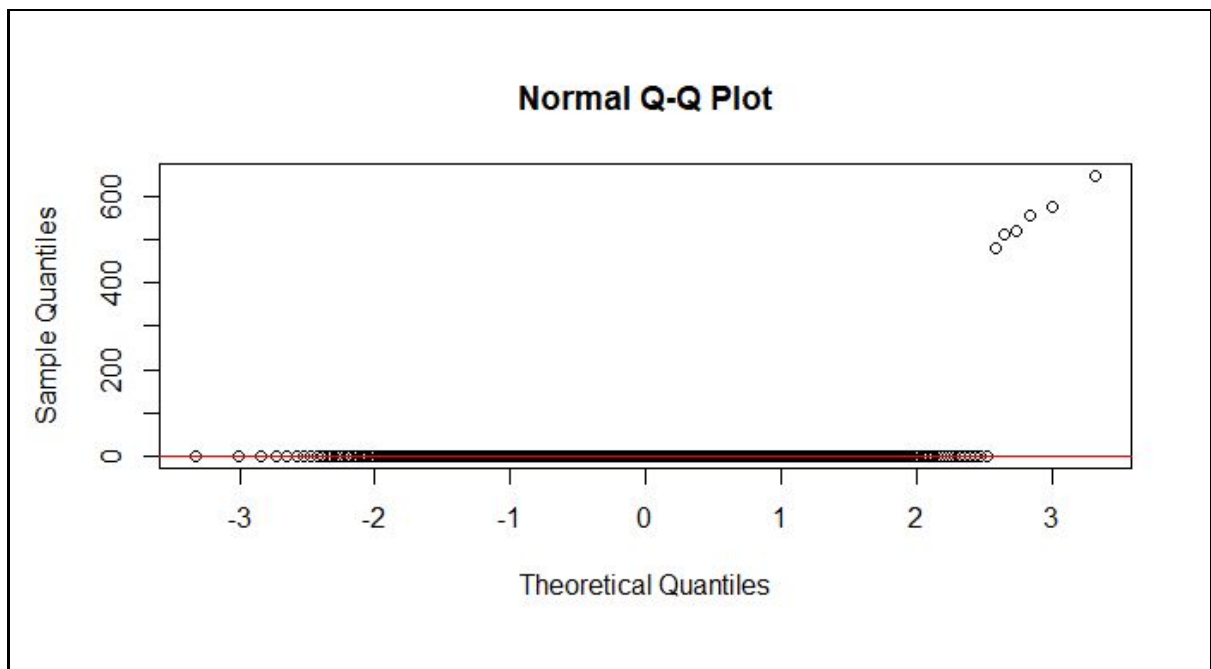
Gráfica contra precio de venta



- **PoolArea:** Área de la piscina en metros cuadrados.

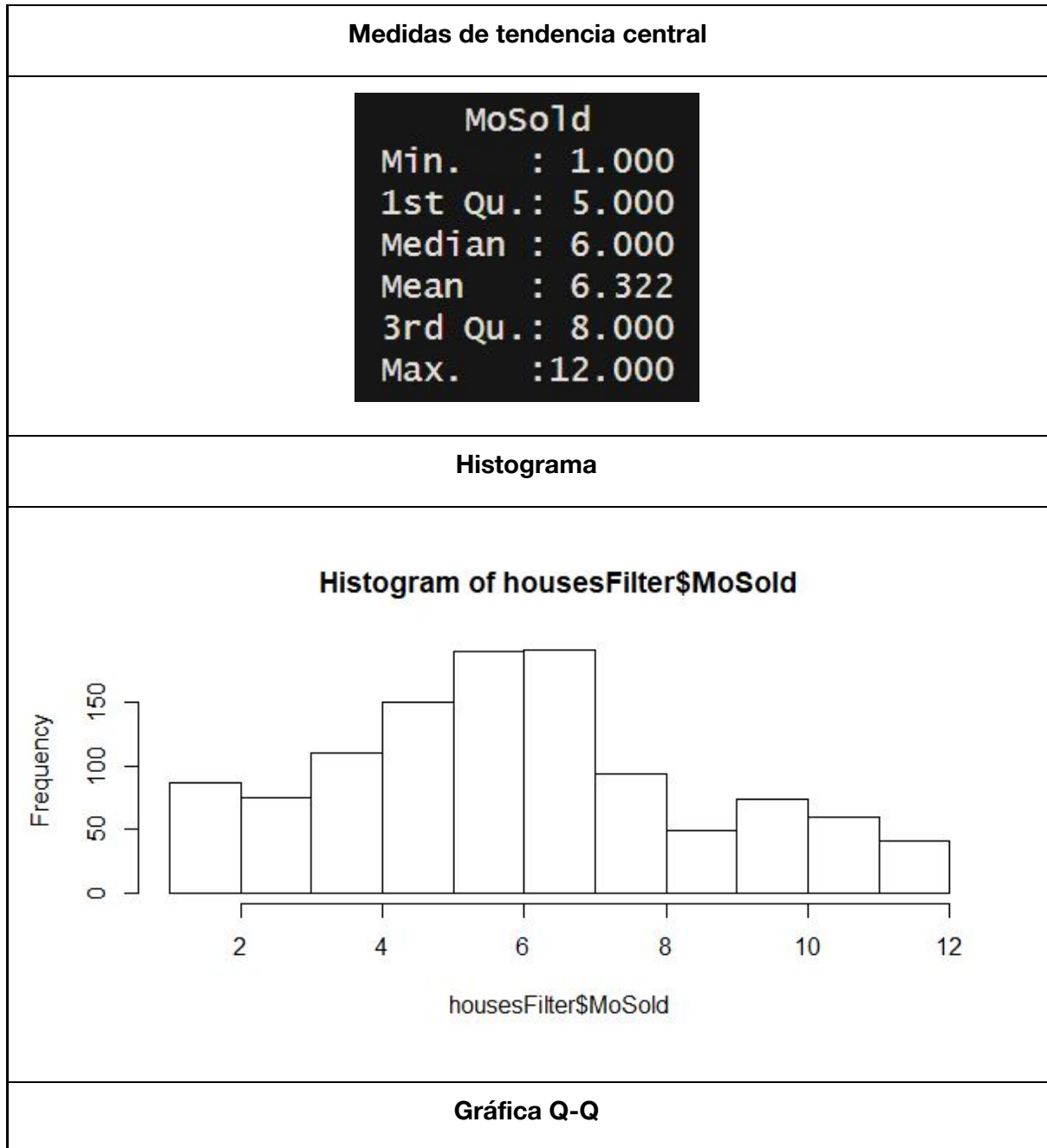
Estos datos no brindan información interesantes para analizar.

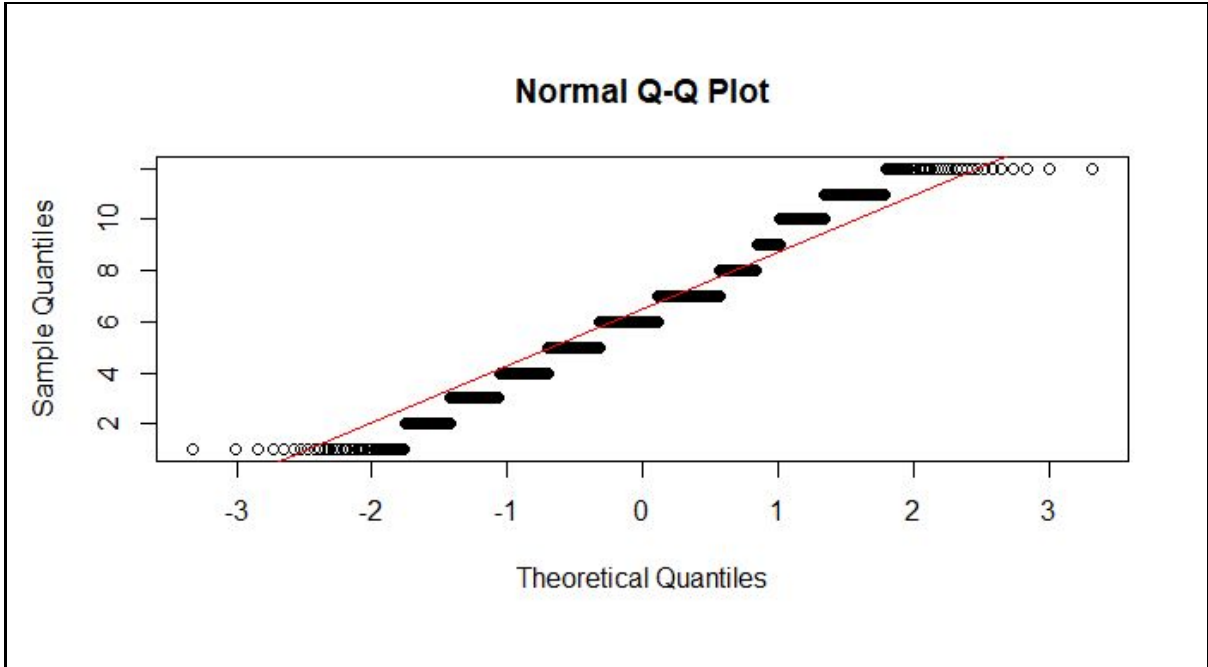




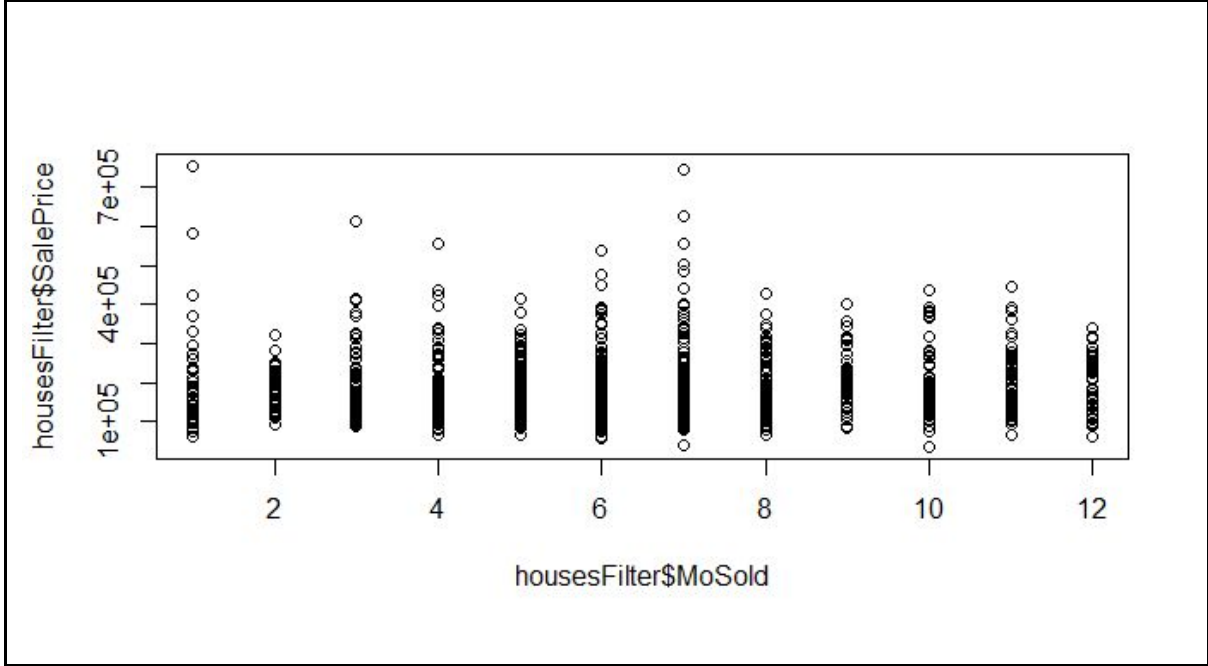
- **MoSold:** Mes en que fue vendida.

La variable MoSold tiene simetría sin embargo por ser datos categóricos no se puede profundizar en su análisis.





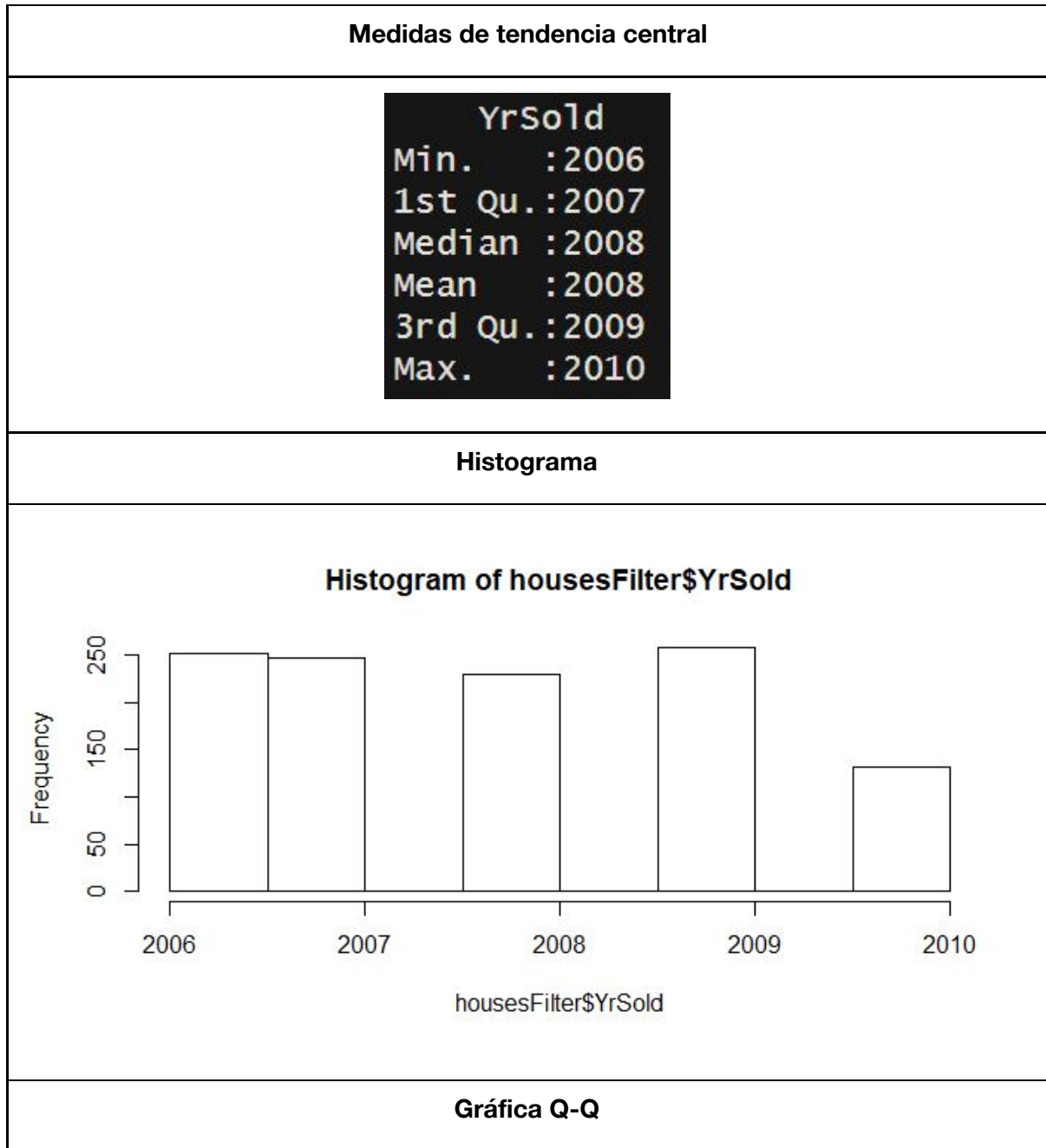
**Gráfica contra precio de venta**

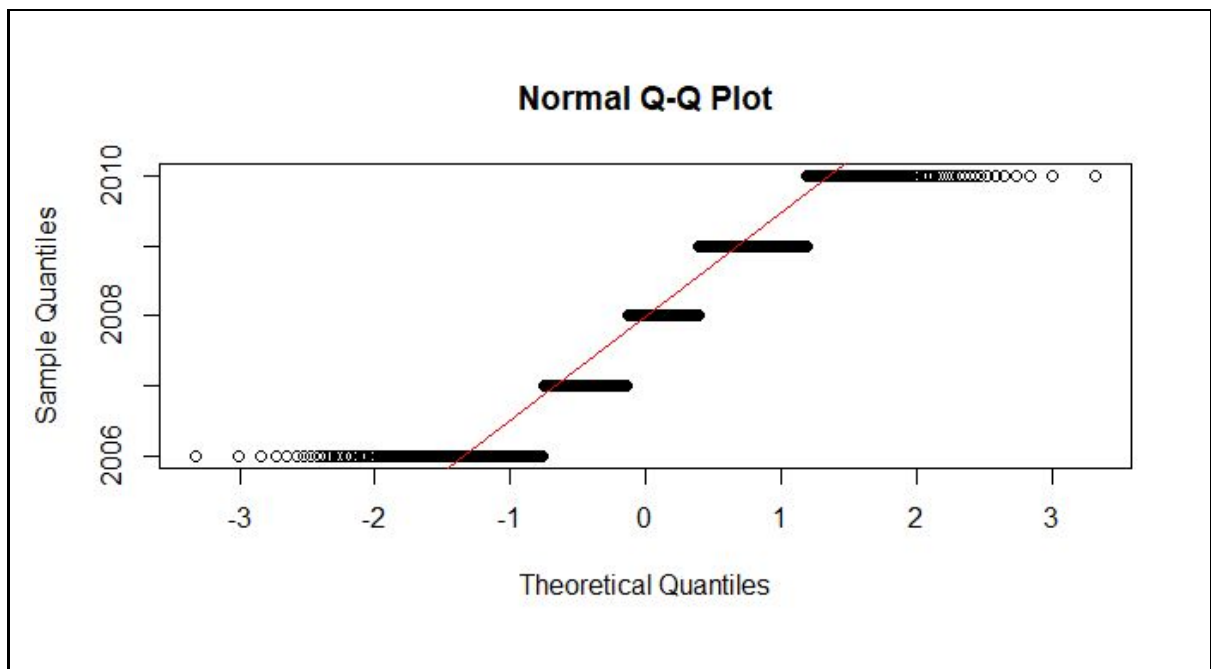




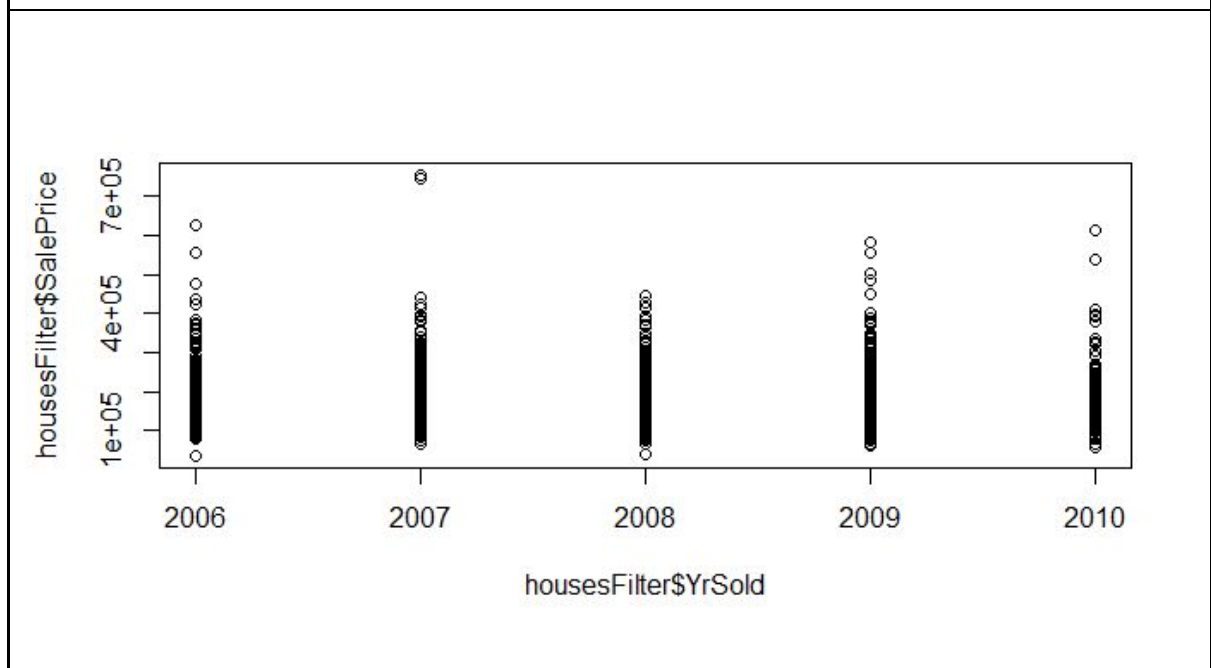
- **YrSold:** Año en que fue vendida.

La variable MoSold no tiene simetría, y por ser datos categóricos no se puede profundizar en su análisis.



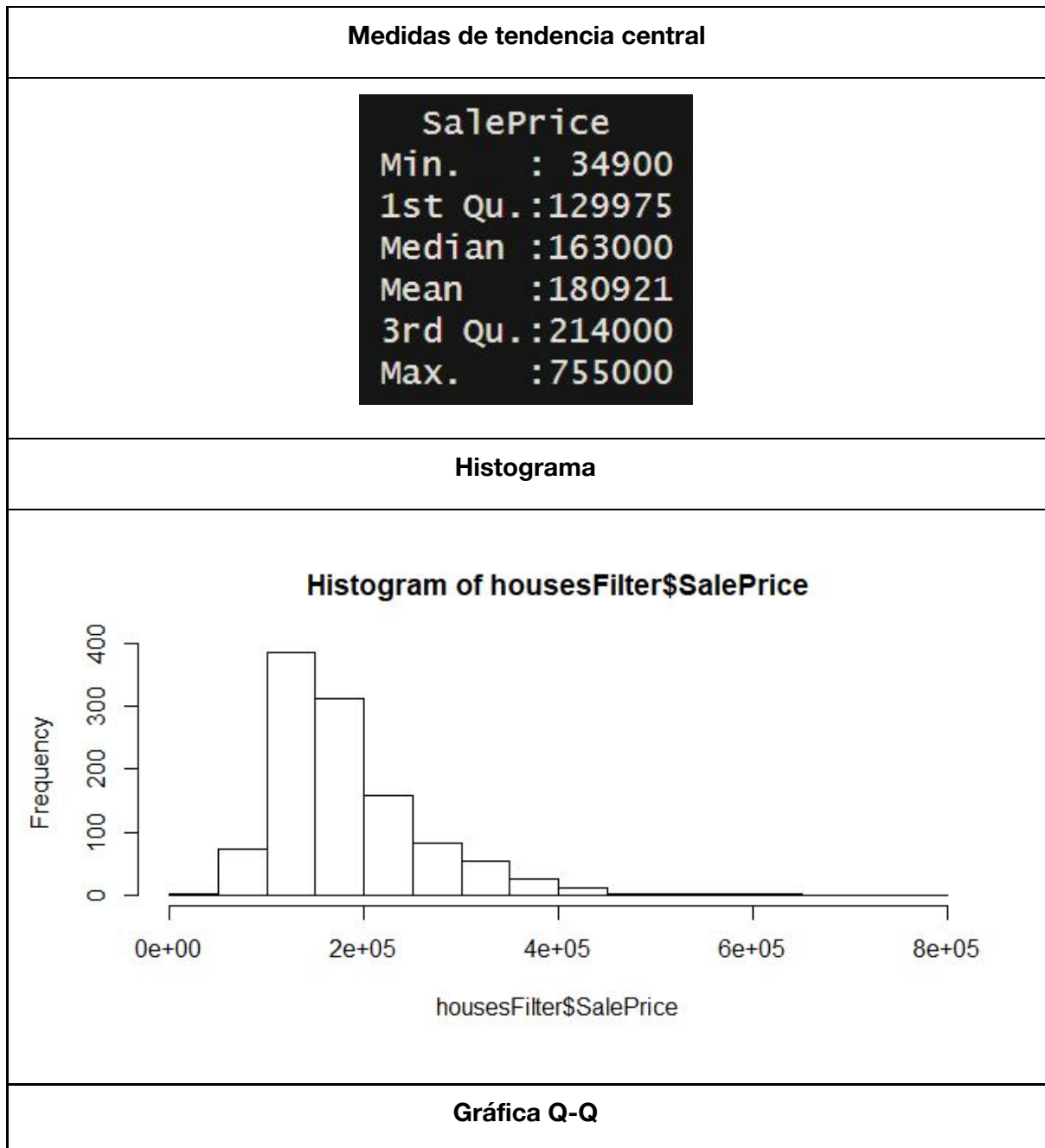


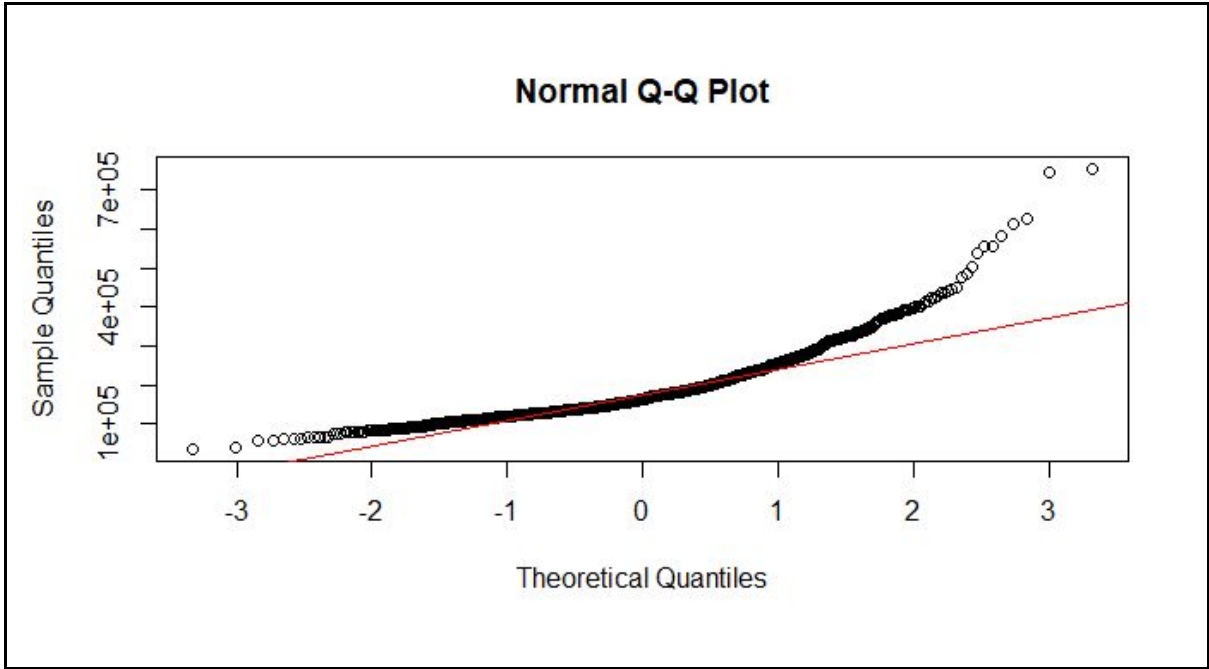
Gráfica contra precio de venta



- **PriceSold:** Precio de venta.

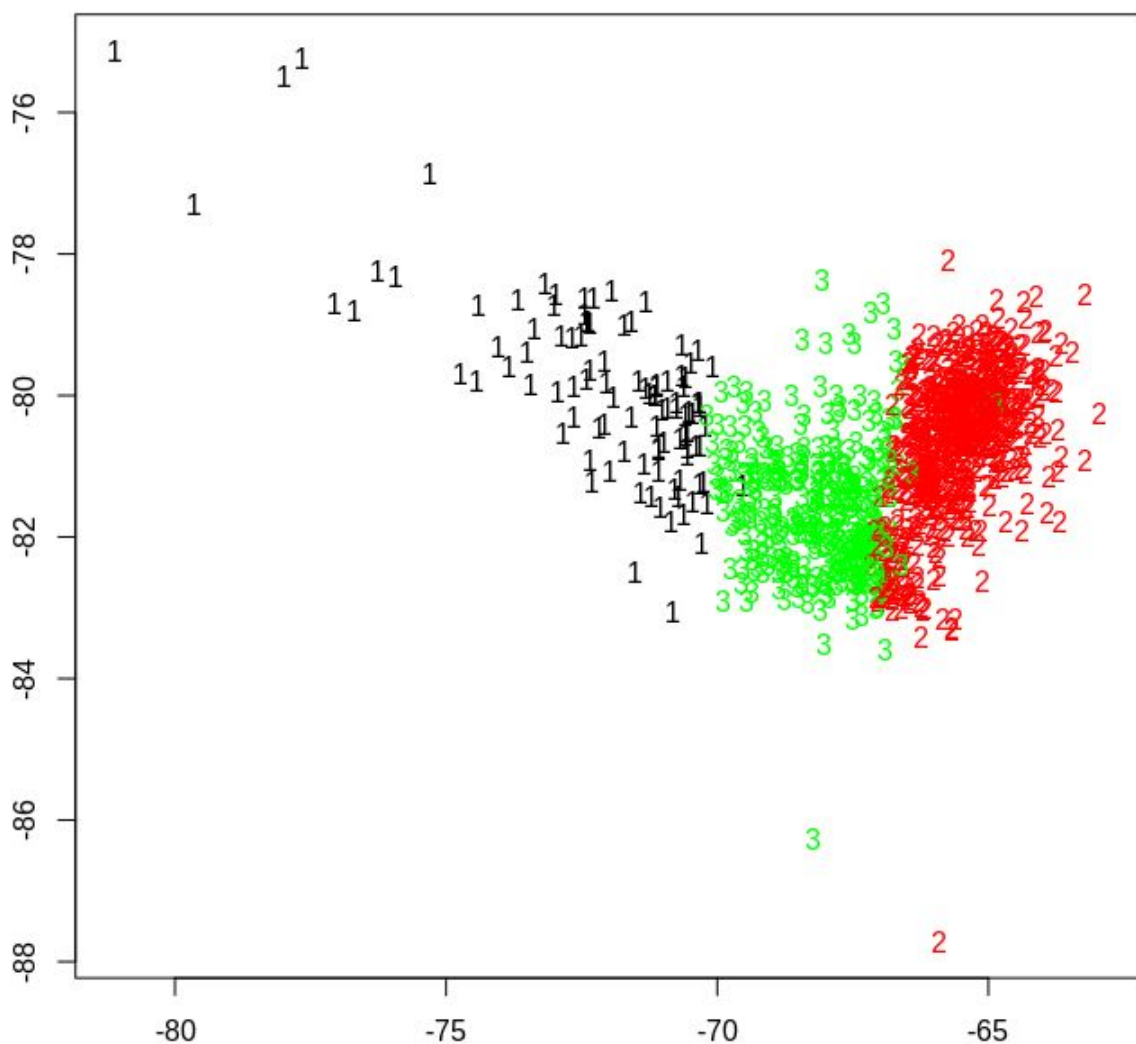
La variable MoSold tiene sesgo positivo y tiene un poco de normalización en sus datos.





## 2. Clustering:

En el caso del análisis de grupos se inició, tomando las variables previamente descritas y realizando un clúster con el algoritmo de Kmeans. Se consideran las variables anteriores , como aquellas que influyen en el precio de una casa, sin embargo el valor de silueta de 0.561677. A continuación se muestra el clúster y las medidas de tendencia central para cada uno de los clústeres obtenidos.



**Figura No. 1** Representación gráfica cluster No.1

SalePrice	
Min.	:301000
1st Qu.	:320000
Median	:348000
Mean	:377265
3rd Qu.	:395000
Max.	:755000

**Figura No. 2** Medidas de tendencia central para el precio de venta grupo No. 1

SalePrice	
Min.	: 35311
1st Qu.	:118450
Median	:135000
Mean	:134136
3rd Qu.	:154400
Max.	:178000

**Figura No. 3** Medidas de tendencia central para el precio de venta grupo No. 2

SalePrice	
Min.	:178900
1st Qu.	:193000
Median	:216837
Mean	:222938
3rd Qu.	:248900
Max.	:297000

**Figura No. 4** Medidas de tendencia central para el precio de venta grupo No. 3

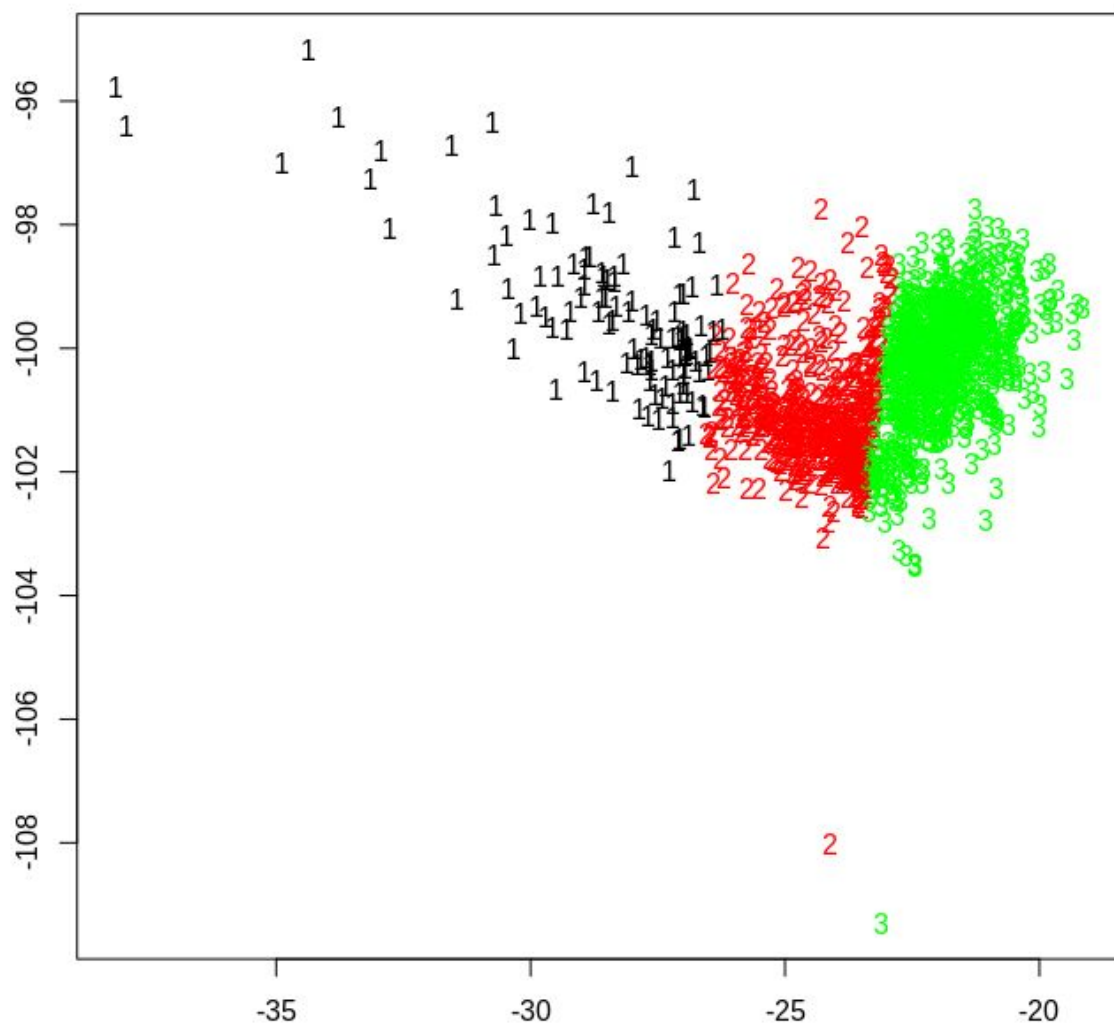
Como se puede observar en las Figuras 2-4. La generación de clústeres induce una clasificación por precio dentro de los grupos en donde *el grupo 2 representa las casas más baratas, el grupo 3 las casas de precio medio y el grupo 1 las casas de mayor precio.*

Gracias a que el valor de la silueta se encuentra dentro de los límites para considerarse aceptable, se optó por evaluar los valores de correlación entre las diferentes variables y el

precio. A continuación se presenta cada una de las variables con su valor de correlación correspondiente.

LotFrontage	0.3442698
LotArea	0.2999622
YearBuilt	0.5253936
YearRemodAdd	0.5212533
MasVnrArea	0.4886582
BsmtFinSF1	0.3903005
BsmtFinSF2	-0.02802137
BsmtUnfSF	0.2131287
TotalBsmtSF	0.6156122
X1stFlrSF	0.6079691
X2ndFlrSF	0.306879
LowQualFinSF	-0.001481983
GrLivArea	0.7051536
TotRmsAbvGrd	0.5470674
Fireplaces	0.4618727
GarageYrBlt	0.504753
GarageCars	0.6470336
GarageArea	0.6193296
WoodDeckSF	0.336855
OpenPorchSF	0.3433538
EnclosedPorch	-0.1548432
ScreenPorch	0.1104268
PoolArea	0.09248812
MoSold	0.05156806
YrSold	-0.01186882

Tomando en cuenta que ninguna presentó una correlación alta , se eliminó toda aquella variable que cuenta con un valor de correlación menor a 0.5. Estas corresponden con aquellas variables que más de un tercio de sus valores son 0, o son de tipo categórico. El clúster obtenido se muestra a continuación, con las medidas de tendencia central para cada uno de los grupos.



**Figura No. 5** Representación gráfica cluster No.2

SalePrice	
Min.	:172400
1st Qu.:	185888
Median	:208400
Mean	:215390
3rd Qu.:	239546
Max.	:294000

**Figura No. 6** Medidas de tendencia central para el precio de venta grupo No. 1, clúster 2.



SalePrice	
Min.	:295000
1st Qu.:	318000
Median	:342643
Mean	:373287
3rd Qu.:	394808
Max.	:755000

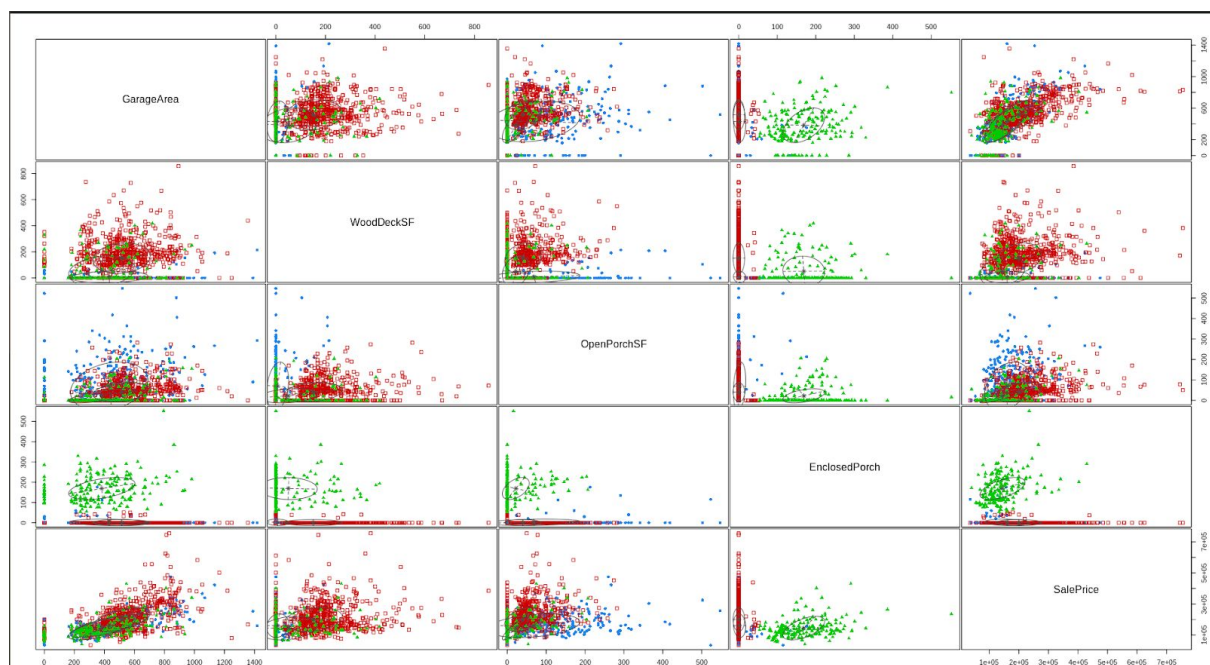
**Figura No. 7** Medidas de tendencia central para el precio de venta grupo No. 2,clúster 2.

SalePrice	
Min.	: 34900
1st Qu.:	112500
Median	:133000
Mean	:129105
3rd Qu.:	148500
Max.	:172000

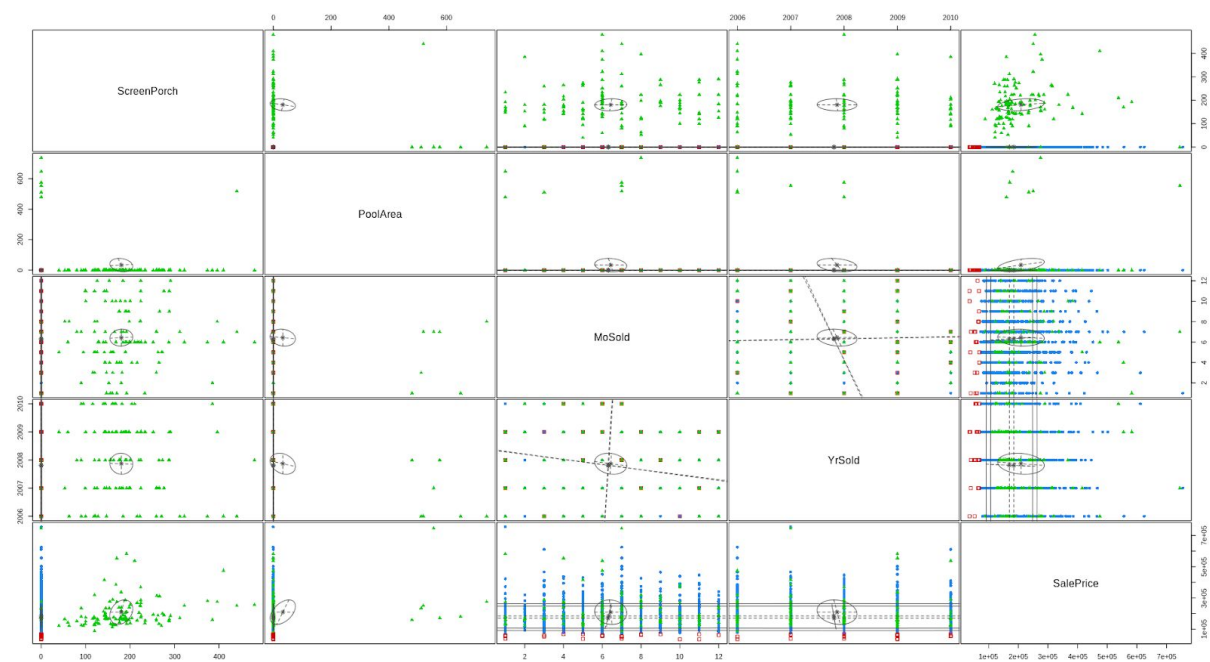
**Figura No. 8** Medidas de tendencia central para el precio de venta grupo No. 3, clúster 2.

Sin embargo el valor de silueta disminuye de 0.561677, a 0.5612973. Lo que disminuye la distancia entre clústeres, en específico , el clúster No.1 y No.2. A medida que se repite el proceso , la silueta del clúster no mejora por lo que se descarta del análisis de grupos.

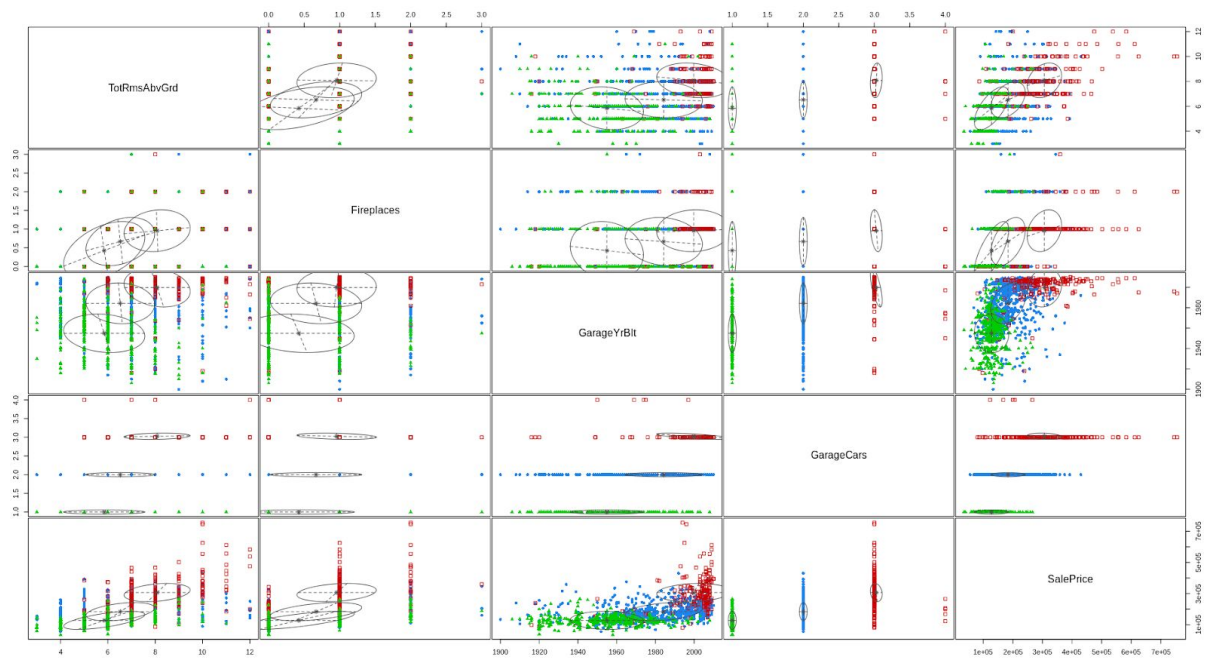
Con el objetivo de observar la manera en que se comportan los agrupamientos de las diferentes variables en relación al precio de venta se realizó un análisis de agrupamiento utilizando Mixture of gaussians, a continuación se presentan los resultados obtenidos.



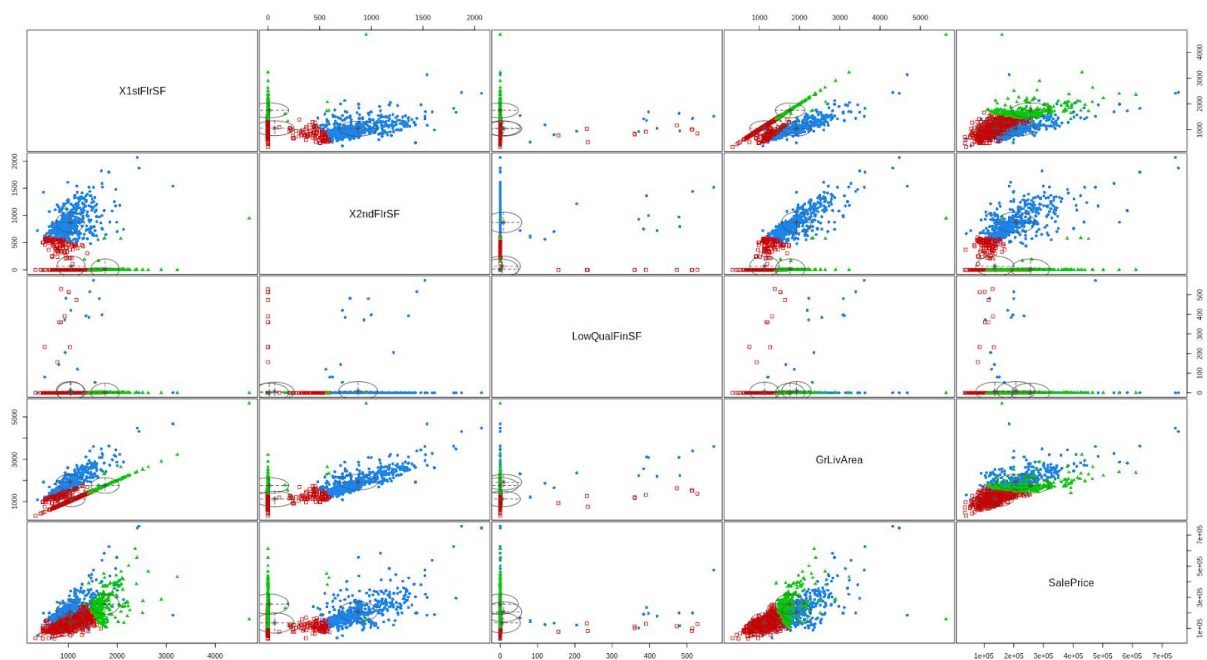
**Figura No. 9** Clústeres obtenidos utilizando Mixture of gaussians



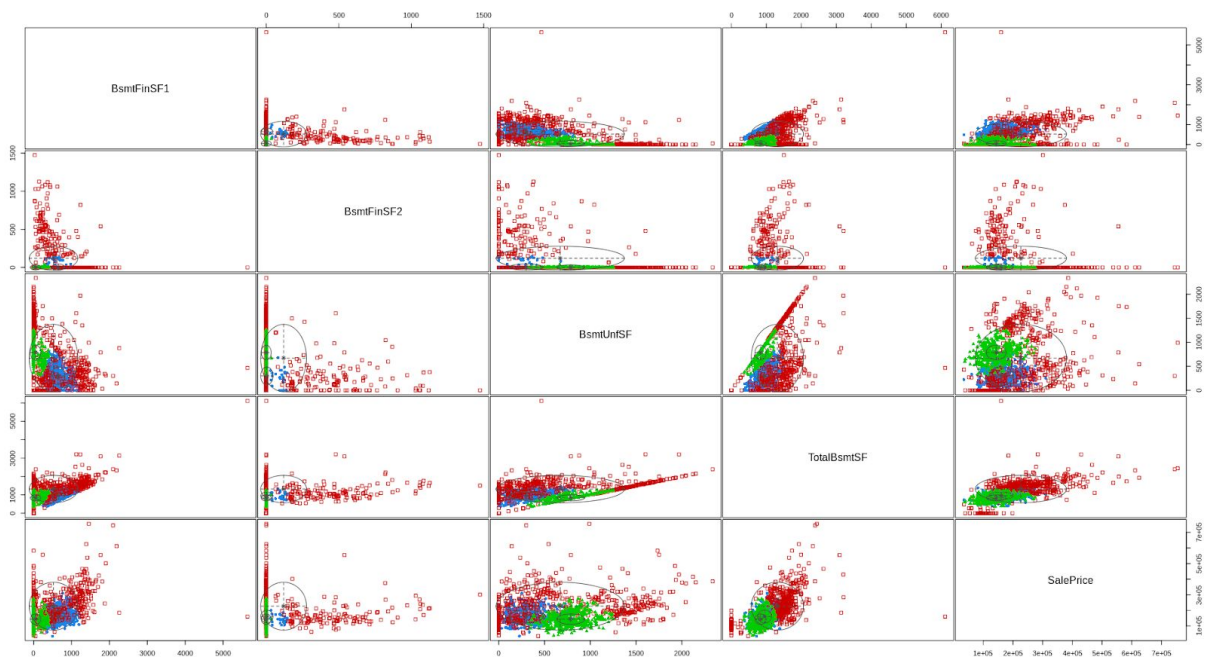
**Figura No. 10** Clústeres obtenidos utilizando Mixture of gaussians



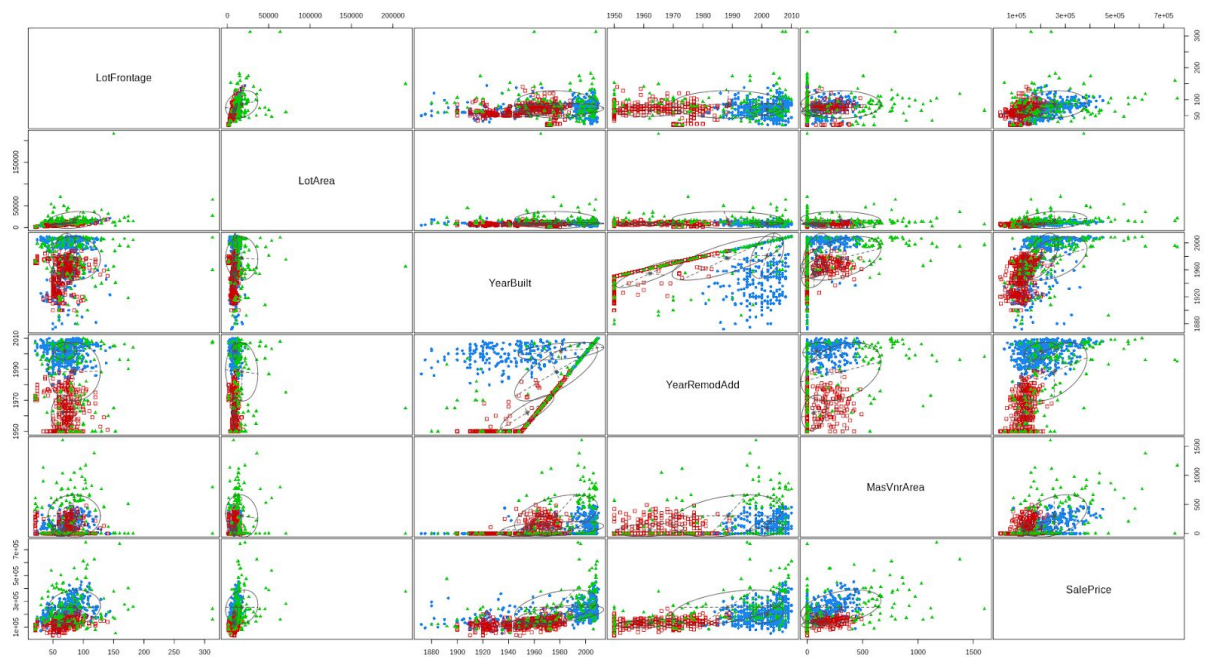
**Figura No. 11** Clústeres obtenidos utilizando Mixture of gaussians



**Figura No. 12** Clústeres obtenidos utilizando Mixture of gaussians



**Figura No. 13** Clústeres obtenidos utilizando Mixture of gaussians



**Figura No. 14** Clústeres obtenidos utilizando Mixture of gaussians

Como se puede observar en la mayoría de gráficas generadas, de las diferentes variables en relación al precio de venta , no se puede identificar de forma verídica 3 clústeres. Tomando únicamente las variables en las que sí es posible diferenciar se obtuvo una silueta de 0.5617, mediante el método de Kmeans y 0.46 mediante mixture of gaussians. Gracias a que este es el valor de silueta más alto obtenido hasta el momento, estas serán las variables a considerar para generar los clusters para generar los árboles.



```

> summary(g1$SalePrice)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
178900 193000 216837 222938 248900 297000
> summary(g2$SalePrice)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 35311 118450 135000 134136 154400 178000
> summary(g3$SalePrice)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
301000 320000 348000 377265 395000 755000

```

En conclusión, las casas están agrupadas de esta forma:

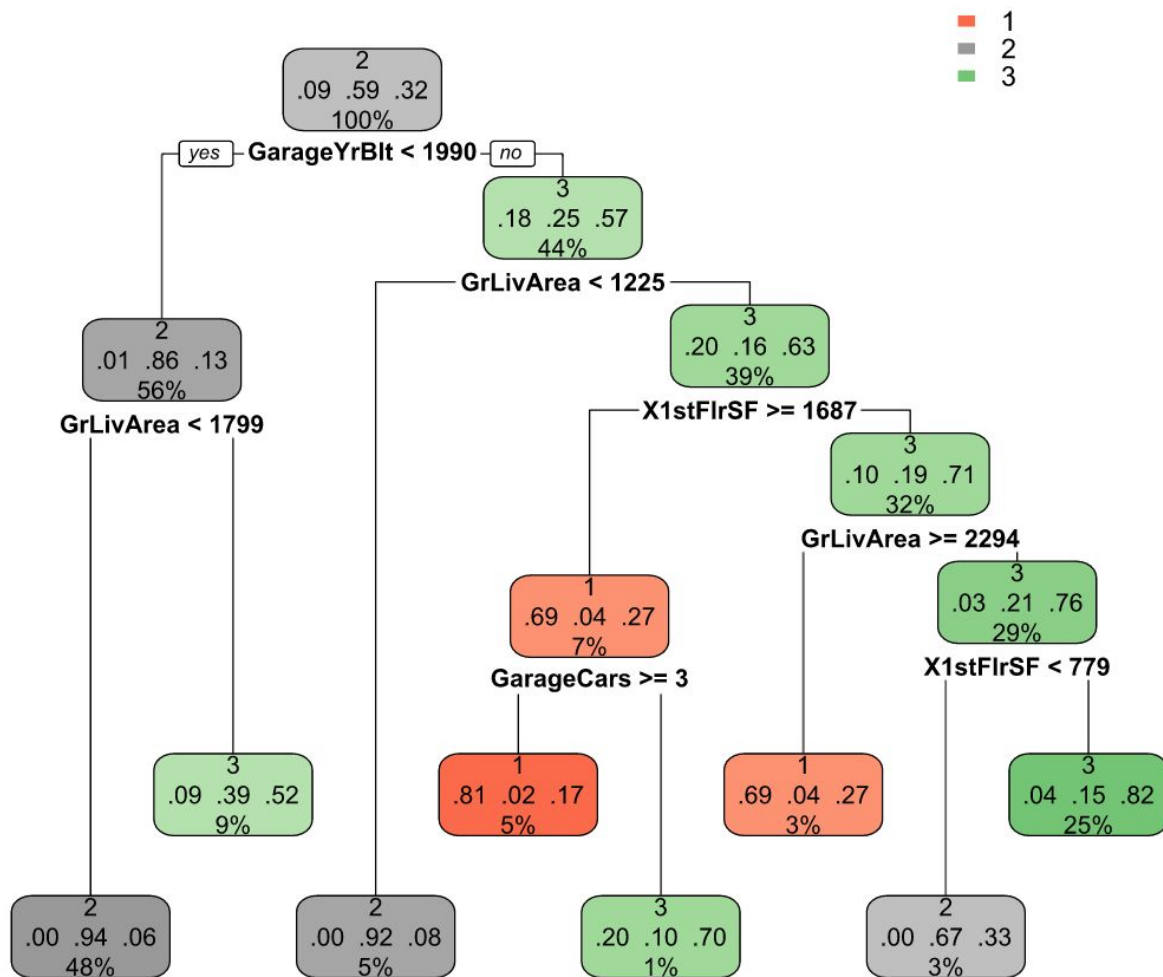
- Económica: \$. 35,311 - \$. 178,000
- Medio: \$. 178,900 - \$. 297,000
- Caro: \$. 301,000 - \$. 755,000

### 3. Árbol de clasificación:

A partir de los análisis previos del análisis exploratorio y de clustering, se decidió realizar un árbol de clasificación con las variables que se muestran a continuación:

- GarageYrBlt
- GrLivArea
- X1stFlrSF
- GarageCars

Se considera que estas son las variables que definen a qué grupo pertenece una casa: económica, media y cara.



**Figura No. 14** Árbol de clasificación.

Del mismo modo, se realizó una matriz de confusión para medir la eficiencia del algoritmo, que se muestra a continuación:

	1	2	3
1	78	21	10
2	21	169	0
3	12	0	26

**Figura No. 15** Matriz de confusión para árbol de clasificación

De la matriz de confusión podemos notar que:

- De **109** casas que pertenecen al grupo 1, **78 casas** se clasificaron correctamente como casas de mayor precio, **21 casas** las clasificó erróneamente como las más baratas y **10** como casas de precio medio, también de forma errónea.

- De **190** casas que pertenecen al grupo 2, **169 casas** se clasificaron correctamente como casas de precio bajo y **21 casas** las clasificó como el grupo de casas caras erróneamente.
- De **38** casas que pertenecen al grupo 3, **26 casas** se clasificaron correctamente como casas de precio medio y **12 casas** las clasificó erróneamente como las más caras
- El algoritmo tiene problemas para distinguir entre casas caras y casas baratas.
- El algoritmo es capaz de distinguir bastante bien entre casas de precio medio con otro tipo de casas.

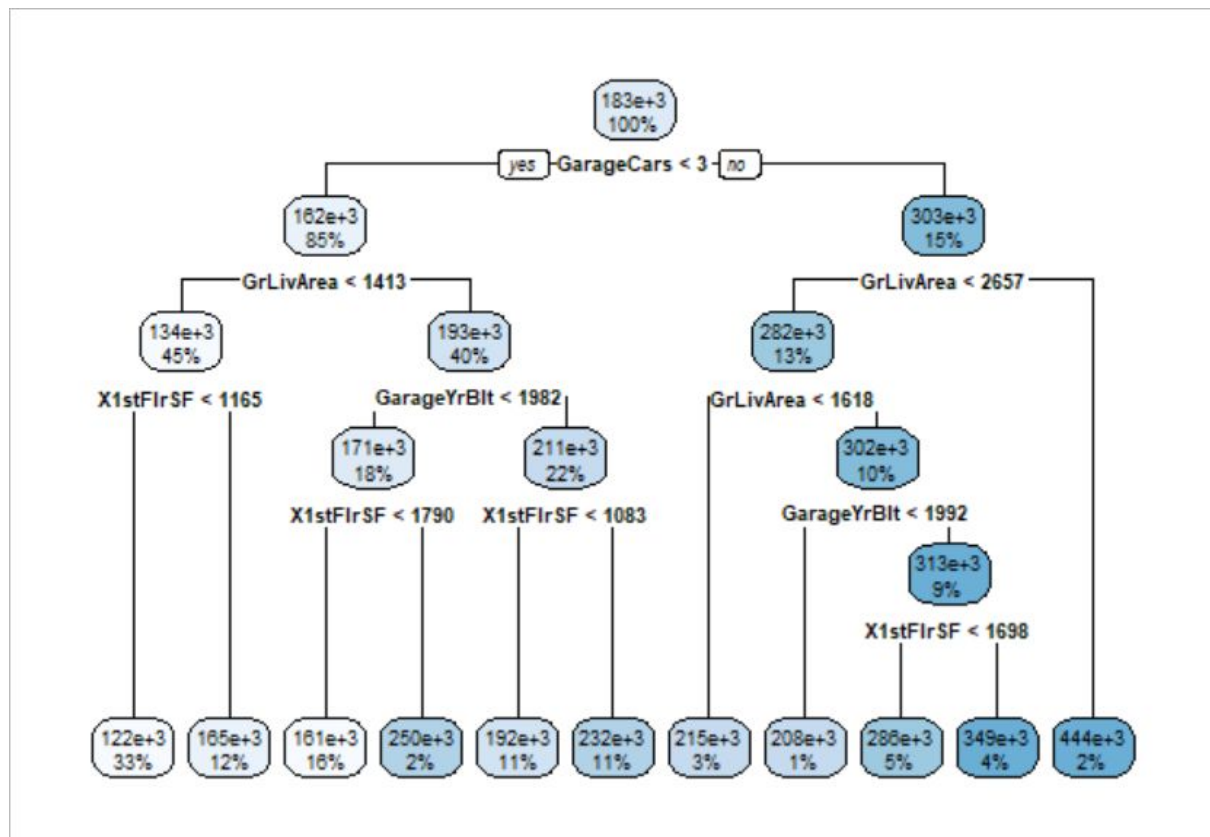
Adicionalmente, se obtuvo un **accuracy de 0.8101**. Lo cuál es un número considerablemente alto por lo que se concluye que sí es un árbol capaz de predecir correctamente.

#### 4. Árbol de regresión:

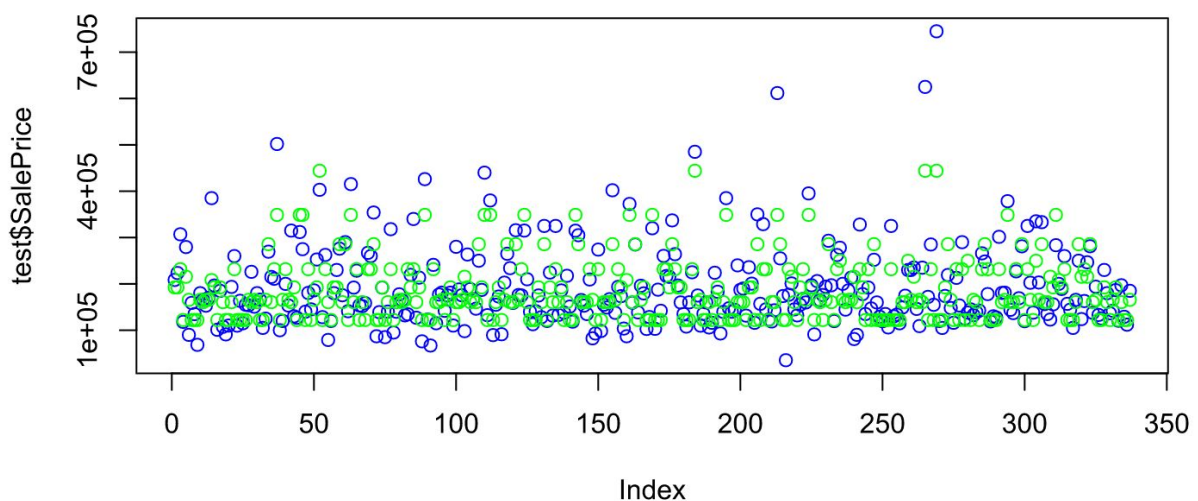
Del mismo modo, con el dataset seleccionado previamente, se decidió realizar un árbol de regresión con las variables que se muestran a continuación:

- |               |              |
|---------------|--------------|
| • GarageYrBlt | • GarageCars |
| • GrLivArea   | • X2stFlrSF  |
| • X1stFlrSF   |              |

Se considera que estas son las variables que definen a qué grupo pertenece una casa: económica, media y cara.



**Figura No. 16** Árbol de regresión.

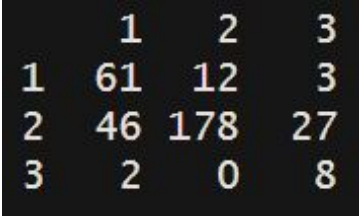


**Figura No. 17** Gráfica de puntos de la predicción y datos reales.

Consideramos que sí logró predecir correctamente el árbol de regresión, pues si vemos la Figura 17, casi todos los puntos están correctos sin embargo los puntos más altos perjudican al modelo, porque también podemos ver que el RMSE es 49,029, lo cuál es muy alto por lo antes mencionado.



## 5. Random forest:



	1	2	3
1	61	12	3
2	46	178	27
3	2	0	8

**Figura No. 18** Matriz de confusión para árbol del random forest.

De la matriz de confusión podemos notar que:

- De **76** casas que pertenecen al grupo 1, **61 casas** se clasificaron correctamente como casas de precio medio, **12 casas** las clasificó erróneamente como las de precio bajo y **3** como casas de precio alto, también de forma errónea.
- De **251** casas que pertenecen al grupo 2, **178 casas** se clasificaron correctamente como casas de precio bajo, **46 casas** las clasificó erróneamente como casas de precio medio y **27 casas** las clasificó como el grupo de casas caras erróneamente.
- De **10** casas que pertenecen al grupo 3, **8 casas** se clasificaron correctamente como casas de precio alto y **2 casas** las clasificó erróneamente como las medio.
- El algoritmo tiene problemas para distinguir entre casas medias y casas baratas.
- El algoritmo es capaz de distinguir bastante bien entre casas de precio alto con otro tipo de casas.

También, se obtuvo un **accuracy de 0.7329**, y aunque este es un número considerablemente alto, se considera mejor utilizar el árbol de regresión o de clasificación, pues, tienen una eficiencia más alta.