



ACTIVIDAD 1 DE ESTADÍSTICA

Autor: Asier Matas



Fecha: 12 Dic 2016 (ahí, ahí, rayando el límite, pufff)

Ficheros Necesarios: house_train.csv

House_test.csv

year.txt (creado para el ejercicio)

Número de versión:  secreto profesional

Contenido

1. Actividad propuesta	2
2. Análisis de datos.....	3
3. Limpieza de datos.....	7
4. Datos de trabajo.....	10
5. Analizar efecto superficie – precio vivienda	11
5.1 Modelo incremento total por pie cuadrado	12
5.2 Modelo incremento porcentual por pie cuadrado	15
5.3 Modelo con estadística robusta.....	19
5.4 Modelo de porcentaje con estadística robusta	21
6. Comparación modelos a datos de test.....	23
7. Modelo predictivo	24
8. Aplicación modelos a datos de test	27
9. Entregables.....	27
8.1 Documento de construcción del modelo.....	27
8.2 Documento de análisis del efecto de la superficie de la vivienda en el precio.....	27
8.3 Fichero Test con columna de precio resultante.....	27



1. Actividad propuesta

Introducción: Habéis sido contratados para realizar un estudio sobre el precio de las viviendas en estados unidos. El proyecto tiene dos objetivos:

- 1.- Analizar el efecto de la superficie de la vivienda en el precio de la vivienda
- 2.- Estimar el precio de venta de unos inmuebles de la cartera de la empresa.

Datos: Para esto os entregan dos ficheros de datos `house_train.csv` y `house_test.csv`. El primer fichero contiene datos de viviendas con su precio, pero el segundo no incluye el precio y es este el conjunto de viviendas a valorar.

Entregables: El objetivo es entregar los siguientes entregables (Fecha límite: 09/03/2016):

- Documento de análisis del efecto de la superficie de la vivienda en el precio. (este análisis se realizará sobre el fichero `house_train.csv`). El análisis deberá basarse en alguna técnica aprendida en la asignatura.
- Fichero `house_test.csv` incluyendo una nueva columna con la estimación del precio.
- Documento de construcción del modelo. El modelo deberá ser alguno de los utilizados en la asignatura.

Todos los scripts de R y ficheros auxiliares necesarios que permitan replicar los análisis realizados y las conclusiones alcanzadas.

Evaluación: La evaluación de la actividad se realizará en función a los conocimientos demostrados y la efectividad de las estimaciones realizadas.


También se ha pedido Sorprender !!!




2. Análisis de datos

Lo primero que vamos a hacer es analizar la consistencia, coherencia y validez de los datos aportado.

Siento ser poco fashion Data Science analist, pero voy a echarles un vistazo a ver como esta los datos tanto desde el punto de vista de R como en el viejo estilo de un pobre consultor funcional: con Excel

Y en ciertos casos, para aclarar ciertas dudas y recabar información necesaria para el análisis de datos también consultaremos al oráculo que todo lo sabe: 

Es decir todas las herramientas a nuestro alcance que nos puedan ayudar a hacer un mejor análisis . 

Primera fuente de información para análisis de datos: el propio enunciado del ejercicio:

Campos: Los campos de los ficheros son los siguientes, aunque hay algunos campos que no se conoce muy bien su significado.

Id: identificador de la vivienda

date: fecha asociada a la información

price: precio de la vivienda

bedrooms: número de habitaciones

bathrooms: número de baños

sqft_living=superficie de la vivienda (en pies)

sqft_lot: superficie de la parcela (en pies)

floors: número de plantas

waterfront: indicador de estancia en primera línea al mar

view: número de orientaciones de la vivienda

condition: campo desconocido

grade: campo desconocido



sqft_above: campo desconocido → este parece ser la superficie en pies cuadrados de lo construido por encima del nivel del suelo

sqft_basement: campo desconocido → este parece ser la superficie en pies cuadrados del sótano

yr_built: año de construcción

yr_renovated: año de reforma

zipcode: código postal

lat: latitud

long: longitud

sqft_living15: campo desconocido

sqft_lot15: campo desconocido

Ya tenemos cierta información:

- Se trata de un país anglo sajón: medición en pies cuadrados. Por lo que tendremos que buscar información suplementaria en páginas de dicho entorno
- Nos ayudaremos del siguiente conversor:
<https://www.google.es/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=conversor%20pies%20cuadrados%20metros%20cuadrados>
- Se distingue superficie vivienda y superficie parcela (jardín etc...)
- Se distingue superficie en sótano y alzada

En paralelo vamos a subir los datos a R y echarles un vistazo en Excel, R y lo que haga falta para obtener la máxima información posible de los datos.

```
house_train=read.csv("house_train.csv")
```

OK, les echamos un vistazo

```
head(house_train)
```

```
> head(house_train)
  id      date  price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition grade
1 7129300520 20141013T000000 221900      3      1.00      1180    5650      1         0      0      3      7
2 6414100192 20141209T000000 538000      3      2.25      2570    7242      2         0      0      3      7
3 5631500400 20150225T000000 180000      2      1.00      770    10000      1         0      0      3      6
4 2487200875 20141209T000000 604000      4      3.00      1960    5000      1         0      0      5      7
5 1954400510 20150218T000000 510000      3      2.00      1680    8080      1         0      0      3      8
6 7237550310 20140512T000000 1225000      4      4.50      5420   101930      1         0      0      3     11
  sqft_above sqft_basement yr_built yr_renovated zipcode    lat    long sqft_living15 sqft_lot15
1      1180           0      1955           0   98178 47.5112 -122.257      1340      5650
2      2170          400      1951           0   98125 47.7210 -122.319      1690      7639
3       770           0      1933           0   98028 47.7379 -122.233      2720      8062
4      1050          910      1965           0   98136 47.5208 -122.393      1360      5000
5      1680           0      1987           0   98074 47.6168 -122.045      1800      7503
6      3890         1530      2001           0   98053 47.6561 -122.005      4760     101930
> |
```



summary (house_train)

```
> summary (house_train)
```

id	date	price	bedrooms	bathrooms	sqft_living
Min. :1.000e+06	20150427T000000: 111	Min. : 75000	Min. : 0.000	Min. : 0.000	Min. : 290
1st Qu.:2.124e+09	20140625T000000: 109	1st Qu.: 320000	1st Qu.: 3.000	1st Qu.:1.750	1st Qu.: 1420
Median :3.893e+09	20140626T000000: 107	Median : 450000	Median : 3.000	Median :2.250	Median : 1910
Mean :4.574e+09	20140708T000000: 107	Mean : 539367	Mean : 3.369	Mean :2.115	Mean : 2080
3rd Qu.:7.304e+09	20140623T000000: 106	3rd Qu.: 640000	3rd Qu.: 4.000	3rd Qu.:2.500	3rd Qu.: 2550
Max. :9.900e+09	20150325T000000: 105	Max. :7700000	Max. :10.000	Max. :8.000	Max. :13540

sqft_lot	sqft_lot15	floors	waterfront	view	condition	grade
Min. : 520	Min. : 651	Min. :1.000	Min. :0.000000	Min. :0.0000	Min. :1.000	Min. : 1.000
1st Qu.: 5050	1st Qu.: 5100	1st Qu.:1.000	1st Qu.:0.000000	1st Qu.:0.0000	1st Qu.:3.000	1st Qu.: 7.000
Median : 7616	Median : 7620	Median :1.500	Median :0.000000	Median :0.0000	Median :3.000	Median : 7.000
Mean : 15092	Mean : 12776	Mean :1.494	Mean :0.007651	Mean :0.2361	Mean :3.411	Mean : 7.655
3rd Qu.: 10665	3rd Qu.: 10065	3rd Qu.:2.000	3rd Qu.:0.000000	3rd Qu.:0.0000	3rd Qu.:4.000	3rd Qu.: 8.000
Max. :1651359	Max. :871200	Max. :3.500	Max. :1.000000	Max. :4.0000	Max. :5.000	Max. :13.000

sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long
Min. : 290	Min. : 0.0	Min. :1900	Min. : 0.00	Min. :98001	Min. :47.16	Min. : -122.5
1st Qu.:1200	1st Qu.: 0.0	1st Qu.:1952	1st Qu.: 0.00	1st Qu.:98033	1st Qu.:47.47	1st Qu.: -122.3
Median :1560	Median : 0.0	Median :1975	Median : 0.00	Median :98065	Median :47.57	Median : -122.2
Mean :1788	Mean : 292.2	Mean :1971	Mean : 83.11	Mean :98078	Mean :47.56	Mean : -122.2
3rd Qu.:2210	3rd Qu.: 560.0	3rd Qu.:1997	3rd Qu.: 0.00	3rd Qu.:98117	3rd Qu.:47.68	3rd Qu.: -122.1
Max. :9410	Max. :4820.0	Max. :2015	Max. :2015.00	Max. :98199	Max. :47.78	Max. : -121.3

sqft_living15	sqft_lot15
Min. : 399	Min. : 651
1st Qu.:1490	1st Qu.: 5100
Median :1840	Median : 7620
Mean :1986	Mean : 12776
3rd Qu.:2360	3rd Qu.: 10065
Max. :6210	Max. :871200

nrow (house_train)

```
> nrow (house_train)
[1] 17384
```

str(house_train)

```
> str(house_train)
'data.frame': 17384 obs. of 21 variables:
 $ id : num 7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...
 $ date : Factor w/ 368 levels "20140502T000000",...: 164 219 287 219 280 11 57 250 336 302 .
 $ price : num 221900 538000 180000 604000 510000 ...
 $ bedrooms : int 3 3 2 4 3 4 3 3 3 ...
 $ bathrooms : num 1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
 $ sqft_living : int 1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
 $ sqft_lot : int 5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
 $ floors : num 1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront : int 0 0 0 0 0 0 0 0 0 0 ...
 $ view : int 0 0 0 0 0 0 0 0 0 0 ...
 $ condition : int 3 3 3 5 3 3 3 3 3 3 ...
 $ grade : int 7 7 6 7 8 11 7 7 7 7 ...
 $ sqft_above : int 1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
 $ sqft_basement : int 0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built : int 1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
 $ yr_renovated : int 0 1991 0 0 0 0 0 0 0 0 ...
 $ zipcode : int 98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
 $ lat : num 47.5 47.7 47.7 47.5 47.6 ...
 $ long : num -122 -122 -122 -122 -122 ...
 $ sqft_living15 : int 1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
 $ sqft_lot15 : int 5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
```

dim(house_train)

nrow(unique(house_train[1]))

```
> nrow (house_train)
[1] 17384
```

```
> nrow(unique(house_train[1]))
[1] 17273
```



Por lo pronto nos llama la atención de esta información:

- Tenemos los datos de 17384 casas
- 0 valores nulos (NA) bieeeennnnnnn 🤔
- PRECIO: mínimo 75.000 y máximo: 7.700.000. Lo importante es que el 3er cuartil se nos queda en 640.000.
Y la mediana es 450000. Eso nos indica que tendremos un grupo de casas que serán outliers de precio alto.
Hay 675 casas por encima del millón. Es un dato posible. Es más, personalmente deseable incluso :oP
- Hay 111 ID repetidos. Vemos que son casas con valoraciones en distintos momentos en el tiempo. 🤔
- Hay 368 fechas distintas. Es un dato factible. No obstante todas las fechas incluyen T000000. Si quisiéramos hacer un análisis por fecha o año tendríamos que eliminar esa parte.
- Las fechas de construcción son factibles (no aparece ninguna anterior a 1900) y las de valoración son buenas.
- Hay 8 casas sin baño. Bueno, vale, es raro pero posible.
- Hay 10 casas sin habitaciones. Esto ya es un dato muy muy raro. Porque entonces no serían casas. 🤔
- hay 5 casas sin habitaciones ni baños y de ellas, varias con más de un piso. 🤔
De hecho una de ellas, la 3918400017, tiene 0 habitaciones, 0 baños y 3 pisos.... 🤔
Madre ! Que cocina debe tener !!! Seguro es la casa de Ferran Adrià 🤔
Ese no es dato factible.
- Increíblemente más de un 30% tienen medio baño.. (es decir 0.5, 1.5, 2.5...).
Y los que tienen .25 de baño? será que una toalla y un peso cuenta como 0.25 baño? 🤔
Pues curiosamente por ahí van los tiros. Buscando parece ser que dependiendo si el baño tiene solamente inodoro o también ducha/baño cuenta como 0.5 o como 1. Entonces consideramos este dato como correcto.
- Hay 5 casas con un sótano de menos de 4 m2... la casa 2724049222 tiene un sótano de 0,93 m2. Ahí no puede meter no ya el coche smart, ni la lavadora... Será para una caja fuerte ??? venga, lo pasamos como factible.
- Por las coordenadas GPS que ponen, estamos hablando de Seattle, EE.UU. (confirmado lo del mundo anglo sajón). Además entendemos que el precio será en dólares.
- Vemos que allí, también hay mini pisos ya que tenemos 5 pisos de menos de 40 m2 es decir menores de 430 pies2., aunque después se ve que tienen una parcela bastante grande, igual son casas de campo.



- He comprobado que todos los códigos postales son de Seattle y alrededores.
- La casa 3277800845 casa es milagrosa: tienen más superficie la vivienda que la parcela donde se asientan (en un solo piso y sin sótano). 65 pies² más.
La casa 9828702895 tiene un medio piso impresionante: tiene 1.5 pisos , superficie vivienda: 2420 (224m²) y superficie de la parcela 520 (48,3 m²) . Pufff Hay 1656 casas con medio piso... Esto tiene que tener explicación:

Buscando en el Oráculo llegué a estas explicaciones:

"This most likely means that it's a cape cod with a couple of bedrooms upstairs. It's 1.5 instead of 2 because the second floor is half the size of the first floor due to the slope of the roof."

1.5 story cape means the exterior walls go up 1.5 storys the second floor walls are 4 feet and then it has a angled ceiling in the bedrooms the house im working on has 3 bedrooms and one bath all decent sized

Por lo que el dato es correcto por lo que tener pisos 1.5 y 2 lo que hace es indicarnos no tanto si es un piso o dos sino si el segundo piso es del mismo tamaño, pies cuadrados habitables que el segundo.

- Creo que el modelo de previsión que voy a hacer ya lo hicieron antes, porque tenemos 10 casas valoradas en el 2014 que fueron construidas en el 2015. Bueno, creo que no me vayan a desvirtuar el modelo ya que entendemos se habrán valorado con los mismos criterios que el resto. Venga, admitidos por ser ellos.
- Hay entradas de latitud y longitud cargadas con distintos formatos.
- Condition: Buscando nuevamente en el oraculohemos visto que son las condiciones en que esta la casa.

3. Limpieza de datos

Acciones de limpiado de datos:

1.- Vamos a sustituir los datos de la columna de date por una del año. En el análisis de hoy no vamos a fijarnos en el mes/día de valoración. Hoy no haremos tan, tan fino. Pensamos que estamos en un momento estable de la economía y que el que ha hecho las valoraciones también

tiene una vida sin sobresaltos y el día y mes no influye en la valoración. 😊 Además la valoración en el mundo inmobiliario no está influido por eventos del pasado, por series temporales, es decir, dependiendo de que pasó en febrero del 2013, y del 2012 no influye en la valoración del 2014 y 2015. Por lo que no utilizaremos un modelo Arima 🙌🏻

Lo que si haremos es añadir una columna con el año de valoración (2014 o 2015) y borrar la columna de año/més/díaT00000 valoración

Para eso nos creamos un CSV con el id & año. Lo subimos





```
year=read.delim("year.txt")
```

```
head(year)
```

Entregable Actividad 1.R* x		year x
	id	Year
1	7129300520	2014
2	6414100192	2014
3	5631500400	2015
4	2487200875	2014
5	1954400510	2015
6	7237550310	2014

```
total <- cbind(house_train,year)
```

```
head(total)
```

```
> head(total)
      id      date price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition
1 7129300520 20141013T000000 221900      3      1.00      1180      5650      1      0      0      3
2 6414100192 20141209T000000 538000      3      2.25      2570      7242      2      0      0      3
3 5631500400 20150225T000000 180000      2      1.00      770      10000      1      0      0      3
4 2487200875 20141209T000000 604000      4      3.00      1960      5000      1      0      0      5
5 1954400510 20150218T000000 510000      3      2.00      1680      8080      1      0      0      3
6 7237550310 20140512T000000 1225000      4      4.50      5420      101930      1      0      0      3
  grade sqft_above sqft_basement yr_built yr_renovated zipcode      lat      long sqft_living15 sqft_lot15
1      7      1180      0      1955      0 98178 47.5112 -122.257      1340      5650
2      7      2170      400      1951      1991 98125 47.7210 -122.319      1690      7639
3      6      770      0      1933      0 98028 47.7379 -122.233      2720      8062
4      7      1050      910      1965      0 98136 47.5208 -122.393      1360      5000
5      8      1680      0      1987      0 98074 47.6168 -122.045      1800      7503
6     11      3890      1530      2001      0 98053 47.6561 -122.005      4760     101930
      id Year      id Year
1 7129300520 2014 7129300520 2014
2 6414100192 2014 6414100192 2014
3 5631500400 2015 5631500400 2015
4 2487200875 2014 2487200875 2014
5 1954400510 2015 1954400510 2015
6 7237550310 2014 7237550310 2014
> |
```

Y borramos la columna de la fecha que tiene T000000 y el segundo id fruto del cbind

```
total <- subset( total, select = -2 )
```

```
total <- subset( total, select = -21 )
```

```
head(total)
```




```
> head(total)
  id    price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition grade sqft_above
1 7129300520 221900      3       1.00      1180    5650      1          0      0          3      7      1180
2 6414100192 538000      3       2.25      2570    7242      2          0      0          3      7      2170
3 5631500400 180000      2       1.00       770   10000      1          0      0          3      6       770
4 2487200875 604000      4       3.00      1960    5000      1          0      0          5      7      1050
5 1954400510 510000      3       2.00      1680    8080      1          0      0          3      8      1680
6 7237550310 1225000     4       4.50      5420   101930     1          0      0          3     11     3890

  sqft_basement yr_built yr_renovated zipcode    lat    long sqft_living15 sqft_lot15 Year
1             0    1955             0  98178 47.5112 -122.257      1340      5650 2014
2             400    1951           1991  98125 47.7210 -122.319      1690      7639 2014
3             0    1933             0  98028 47.7379 -122.233      2720      8062 2015
4             910    1965             0  98136 47.5208 -122.393      1360      5000 2014
5             0    1987             0  98074 47.6168 -122.045      1800      7503 2015
6            1530    2001             0  98053 47.6561 -122.005      4760     101930 2014
> |
```

2. – Hay 10 casas con 0 habitaciones. Vamos a suponer que hablamos de casas, tal y como dice el fichero y enunciado. Nada de oficinas, almacenes etc..... Por eso Borraremos las 10 casas con 0 habitaciones.

```
total2 <- total[!total$bedrooms == "0", ]
```

```
nrow(total2)
```

```
> nrow(total2)
[1] 17374
```

```
nrow(total)
```

```
> nrow(total)
[1] 17384
```

3. - Eliminamos los datos repetidos. Estos si porque no quiero que los mismos registros al tener precios distintos me desvirtúen el modelo.

```
house_train <- total2
```

```
house_train <- house_train[!duplicated(house_train[,c(1,4:21)]),]
```

```
nrow(house_train)
```

```
> nrow(house_train)
[1] 17346
> |
```

4. Yo trabajaría aún más los datos hasta conseguir unos datos lo mejor posible (digo posible porque muchas veces esa información no está y el cliente no puede arreglarla). Pero como no tenemos aquí al cliente vamos a continuar con los datos que tenemos: Admitimos pulpo como animal de compañía.



4. Datos de trabajo

Echamos un nuevo vistazo de nuevo a nuestro data set.

```
head(house_train)
```

```
summary(house_train)
```

```
> summary(house_train)
      id      price      bedrooms      bathrooms      sqft_living      sqft_lot
Min. :1.000e+06 Min. : 75000 Min. : 1.000 Min. :0.000 Min. : 380 Min. : 520
1st Qu.:2.124e+09 1st Qu.:320000 1st Qu.: 3.000 1st Qu.:1.750 1st Qu.:1423 1st Qu.:5048
Median :3.891e+09 Median :450000 Median : 3.000 Median :2.250 Median :1910 Median :7614
Mean :4.574e+09 Mean :539665 Mean : 3.372 Mean :2.116 Mean :2081 Mean :15095
3rd Qu.:7.304e+09 3rd Qu.:640000 3rd Qu.: 4.000 3rd Qu.:2.500 3rd Qu.:2550 3rd Qu.:10668
Max. :9.900e+09 Max. :7700000 Max. :10.000 Max. :8.000 Max. :13540 Max. :1651359

      floors      waterfront      view      condition      grade      sqft_above
Min. : 1.000 Min. :0.000000 Min. :0.0000 Min. : 1.000 Min. : 3.000 Min. : 380
1st Qu.: 1.000 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.: 3.000 1st Qu.: 7.000 1st Qu.:1200
Median : 1.500 Median :0.000000 Median :0.0000 Median : 3.000 Median : 7.000 Median :1560
Mean : 1.495 Mean :0.007667 Mean :0.2362 Mean : 3.411 Mean : 7.657 Mean :1789
3rd Qu.: 2.000 3rd Qu.:0.000000 3rd Qu.:0.0000 3rd Qu.: 4.000 3rd Qu.: 8.000 3rd Qu.:2210
Max. : 3.500 Max. :1.000000 Max. :4.0000 Max. : 5.000 Max. :13.000 Max. :9410

      sqft_basement      yr_built      yr_renovated      zipcode      lat      long      sqft_living15
Min. : 0.0 Min. :1900 Min. : 0.00 Min. :98001 Min. :47.16 Min. : -122.5 Min. : 399
1st Qu.: 0.0 1st Qu.:1952 1st Qu.: 0.00 1st Qu.:98033 1st Qu.:47.47 1st Qu.: -122.3 1st Qu.:1490
Median : 0.0 Median :1975 Median : 0.00 Median :98065 Median :47.57 Median : -122.2 Median :1840
Mean :292.5 Mean :1971 Mean :83.29 Mean :98078 Mean :47.56 Mean : -122.2 Mean :1986
3rd Qu.:560.0 3rd Qu.:1997 3rd Qu.: 0.00 3rd Qu.:98117 3rd Qu.:47.68 3rd Qu.: -122.1 3rd Qu.:2360
Max. :4820.0 Max. :2015 Max. :2015.00 Max. :98199 Max. :47.78 Max. : -121.3 Max. :6210

      sqft_lot15      Year
Min. : 651 Min. :2014
1st Qu.:5100 1st Qu.:2014
Median :7620 Median :2014
Mean :12773 Mean :2014
3rd Qu.:10070 3rd Qu.:2015
Max. :871200 Max. :2015
```

```
nrow(house_train)
```

ya quitamos los datos que no queriamos

```
> nrow(house_train)
[1] 17346
```

```
str(house_train)
```

```
> str(house_train)
'data.frame': 17346 obs. of 21 variables:
 $ id      : num  7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...
 $ price   : num  221900 538000 180000 604000 510000 ...
 $ bedrooms : int   3 3 2 4 3 4 3 3 3 3 ...
 $ bathrooms : num   1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
 $ sqft_living : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
 $ sqft_lot : int   5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
 $ floors   : num   1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront : int   0 0 0 0 0 0 0 0 0 0 ...
 $ view     : int   0 0 0 0 0 0 0 0 0 0 ...
 $ condition : int   3 3 3 5 3 3 3 3 3 3 ...
 $ grade    : int   7 7 6 7 8 11 7 7 7 7 ...
 $ sqft_above : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
 $ sqft_basement : int   0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
 $ yr_renovated : int   0 1991 0 0 0 0 0 0 0 0 ...
 $ zipcode  : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
 $ lat      : num   47.5 47.7 47.7 47.5 47.6 ...
 $ long     : num  -122 -122 -122 -122 -122 ...
 $ sqft_living15 : int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
 $ sqft_lot15 : int   5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
 $ Year     : int   2014 2014 2015 2014 2015 2014 2014 2015 2015 2015 ...
```

ya nos ha cargado mejor los datos.

```
dim(house_train)
```

Tras la fase de análisis y limpieza de datos. Vamos a por lo que nos piden.



5. Analizar efecto superficie – precio vivienda

En el enunciado nos piden analizar el efecto de la superficie de la vivienda en el precio de la vivienda.

Por eso vamos a hacer justamente un modelo como nos pide exactamente el cliente:



Para ello lo que vamos a hacer es un chequeo muy muy sencillo. Vamos a generar un data frame con la superficie de la vivienda y el precio. Que es exactamente lo que nos piden, aunque no lo necesario para buen análisis.

```
precio_m2 <- house_train[,c(5, 2)]
```

```
head(precio_m2)
```

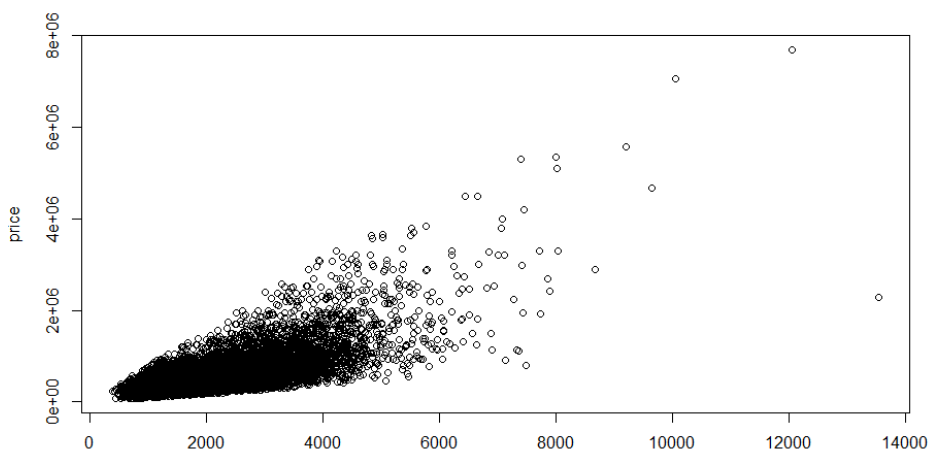
```
> head(precio_m2)
  sqft_living  price
1       1180 221900
2       2570 538000
3        770 180000
4       1960 604000
5       1680 510000
6       5420 1225000
```

```
summary(precio_m2)
```

```
> summary(precio_m2)
  sqft_living      price
Min.   : 380      Min.   : 75000
1st Qu.: 1423     1st Qu.: 320000
Median : 1910     Median : 450000
Mean   : 2081     Mean   : 539665
3rd Qu.: 2550     3rd Qu.: 640000
Max.   :13540     Max.   :7700000
```

OK, vamos a dibujarlo

```
plot(precio_m2)
```





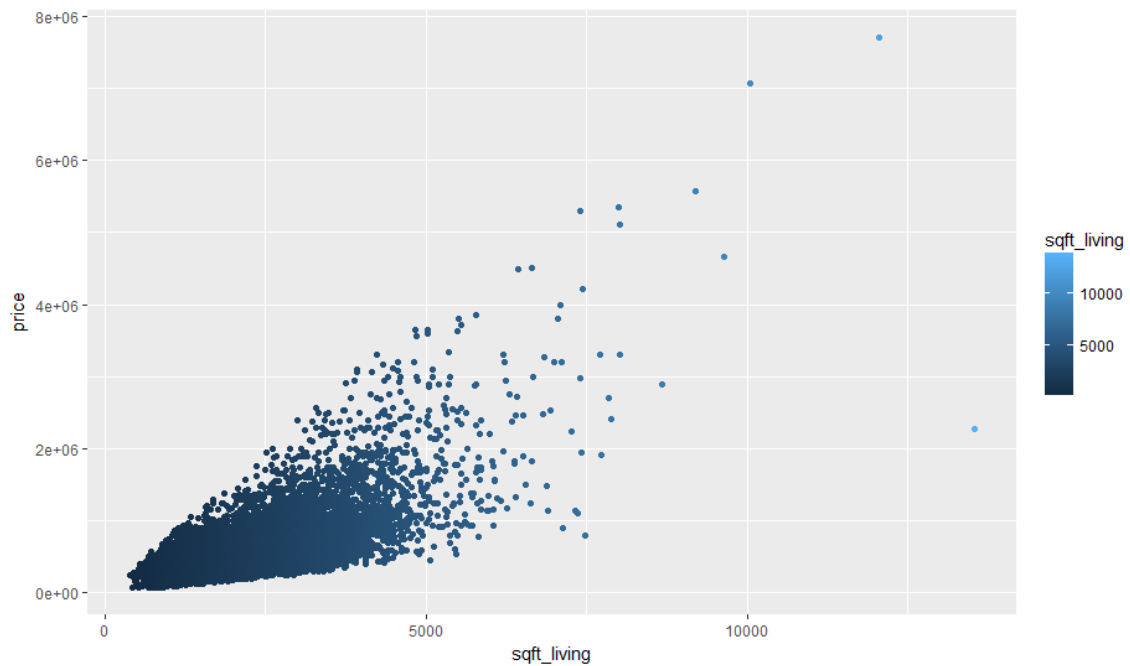
Y ahora en bonito :oP

```
library(ggplot2)
```

```
ggplot(precio_m2)
```

```
G <- ggplot(precio_m2, aes(sqft_living, price))
```

```
G + geom_point(aes(color=sqft_living))
```



Bueno, vamos a hacer el modelo de regresión lineal del precio – superficie habitable de la casa:

5.1 Modelo incremento total por pie cuadrado

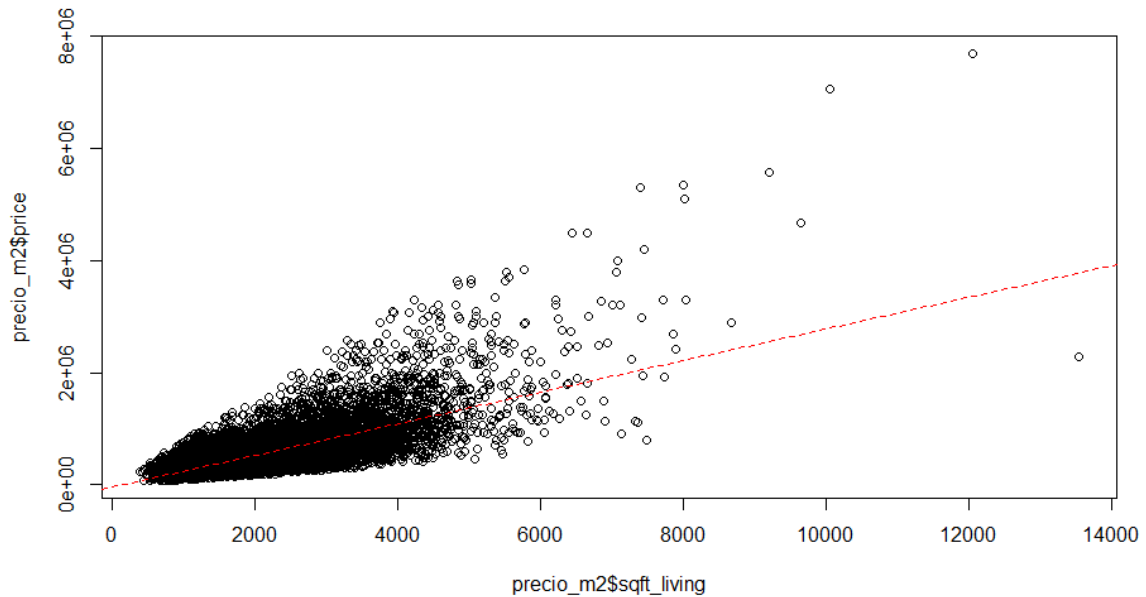
bueno, vamos a ver primero el modelo que nos piden :

```
mod1Precio_m2=lm(price~sqft_living,data=precio_m2)
```

Y lo pintamos:

```
plot(precio_m2$sqft_living, precio_m2$price)
```

```
abline(mod1Precio_m2,col="red",lty = "dashed")
```



Vamos a ver las características del modelo que hemos hecho

`summary(mod1_precio_m2)`

```
Call:
lm(formula = price ~ sqft_living, data = precio_m2)

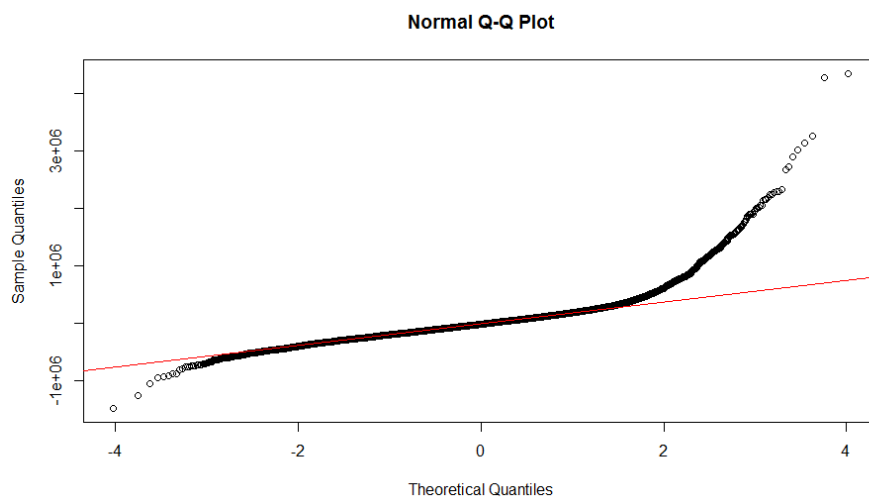
Residuals:
    Min       1Q   Median       3Q      Max
-1491719 -148394  -23715  105786  4348545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -47321.059    4935.186   -9.589  <2e-16 ***
sqft_living   282.056       2.168  130.078  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 263200 on 17344 degrees of freedom
Multiple R-squared:  0.4938,    Adjusted R-squared:  0.4938
F-statistic: 1.692e+04 on 1 and 17344 DF,  p-value: < 2.2e-16
```

Vamos a ver la representación de la relación de la distribución de los residuos versus la distribución normal teórica:

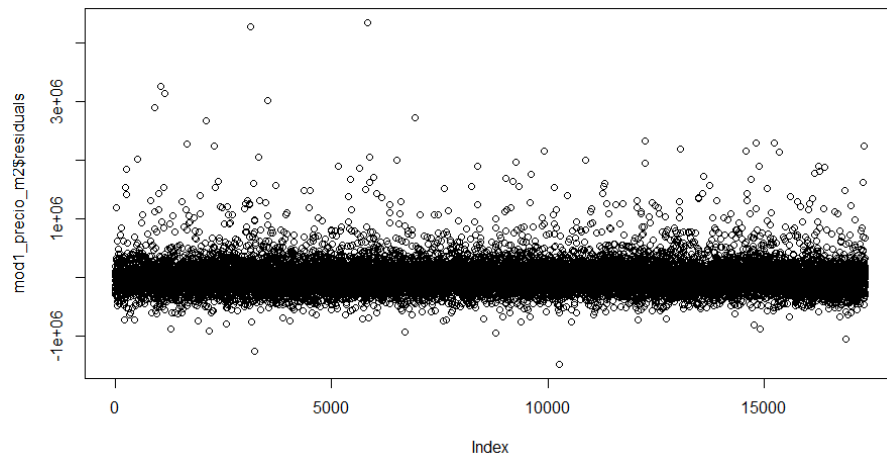
`qqnorm(mod1_precio_m2$residuals); qqline(mod1_precio_m2$residuals,col=2)`



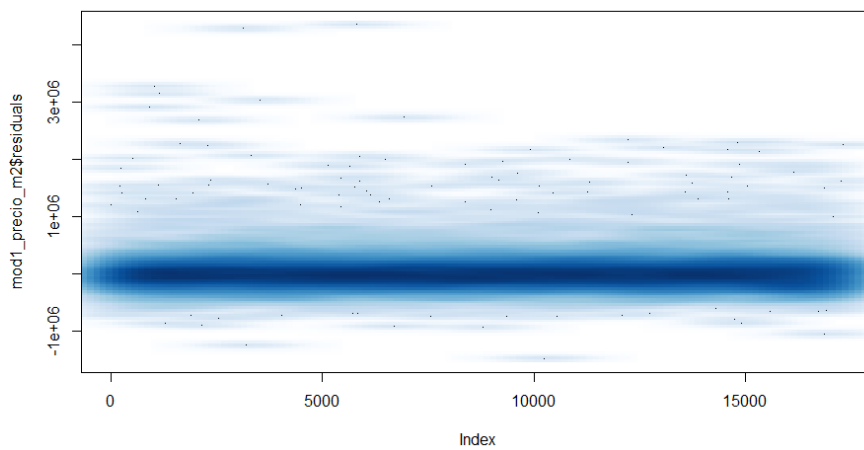


Ahora pintaremos los residuos:

`plot(mod1_precio_m2$residuals)`

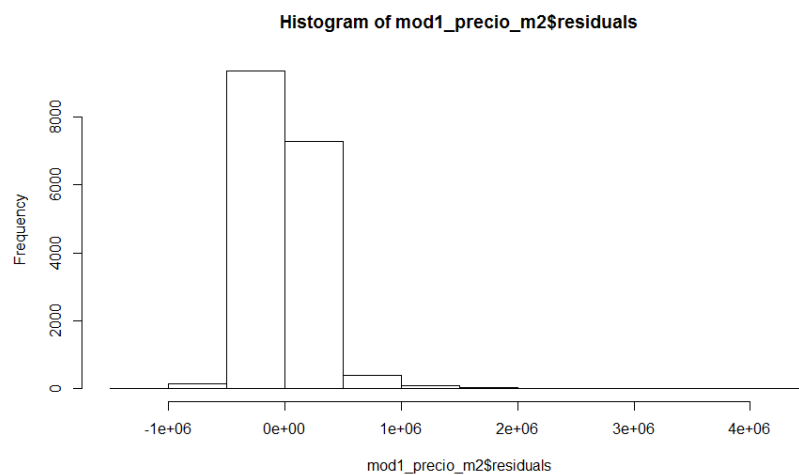


`smoothScatter(mod1_precio_m2$residuals)`



Y el histograma de la distribución de los residuos:

`hist(mod1_precio_m2$residuals)`





Vamos a ver el intervalo de confianza:

```
confint(mod1_precio_m2,level=0.95)
> confint(mod1_precio_m2,level=0.95)
                2.5 %      97.5 %
(Intercept) -56994.5219 -37647.5969
sqft_living   277.8059   286.3063
```

Modelo	Variable Dependiente	Variable Independiente	Interpretación
Regresión Level-Level $y = \beta_0 + \beta_1 x + \epsilon$	y	x	Un aumento de 1 unidad en x se corresponde con un aumento de beta unidades en y. (efecto marginal)

Vamos a hacer un prueba:

```
-47824.434+(282.132*2) - (-47824.434+(282.132*1))
> -47824.434+(282.132*2) - (-47824.434+(282.132*1))
[1] 282.132
```

En este modelo nos sale que, según las observaciones un aumento de 1 pie cuadrado está asociado con un aumento del precio en 282.132 dólares, teniendo en cuenta que hay un margen de error entre - 4.2501 y + 4.2503 (entre 277.8059 y 286.306)

5.2 Modelo incremento porcentual por pie cuadrado

Dado que no estamos muy convencidos vamos a hacer un modelo por porcentual

```
mod2_precio_m2=lm(log(price)~sqft_living,data=precio_m2)
summary(mod2_precio_m2)
```

```
Call:
lm(formula = log(price) ~ sqft_living, data = precio_m2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.95430 -0.28861  0.01527  0.26158  1.27982

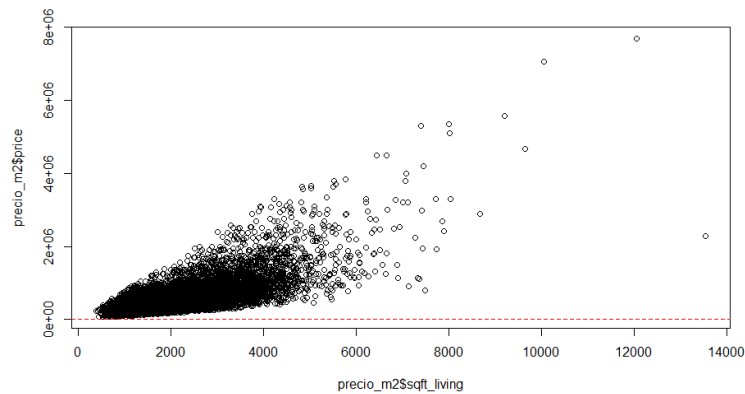
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.222e+01  7.105e-03  1719.9  <2e-16 ***
sqft_living  3.969e-04  3.122e-06   127.1  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3789 on 17344 degrees of freedom
Multiple R-squared:  0.4823,    Adjusted R-squared:  0.4823
F-statistic: 1.616e+04 on 1 and 17344 DF,  p-value: < 2.2e-16
```



Veamos los mismos parámetros que el anterior y comparemos:

```
plot(precio_m2$sqft_living,precio_m2$price)
abline(mod2_precio_m2,col="red",lty = "dashed")
```

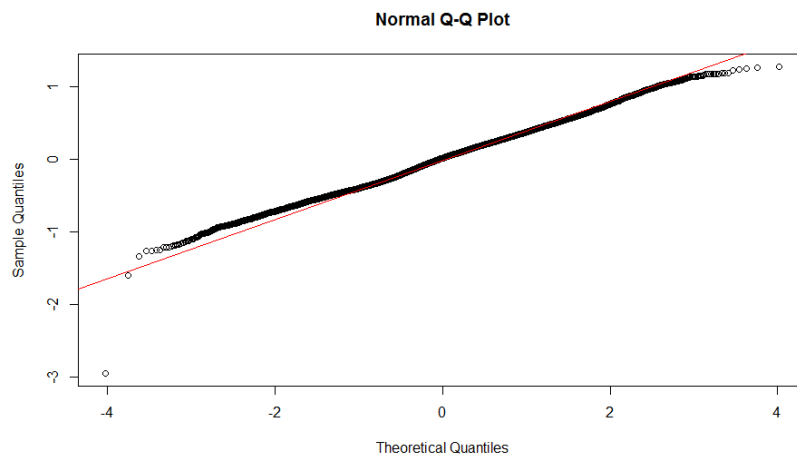


La línea en rojo esta tan abajo porque va en porcentaje y la escala es cientos de miles...

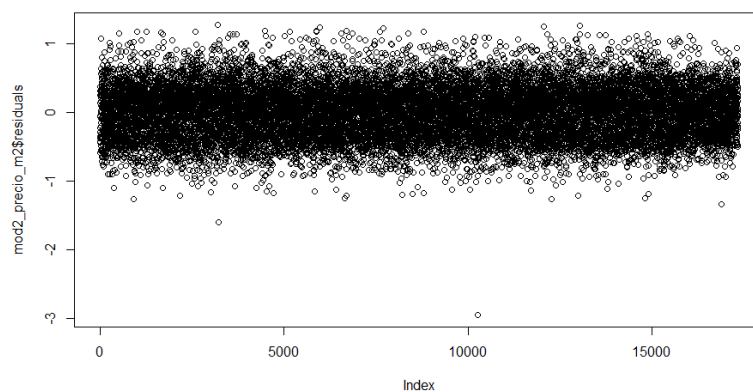


Miramos los residuos:

```
qqnorm(mod2_precio_m2$residuals); qqline(mod2_precio_m2$residuals,col=2)
```

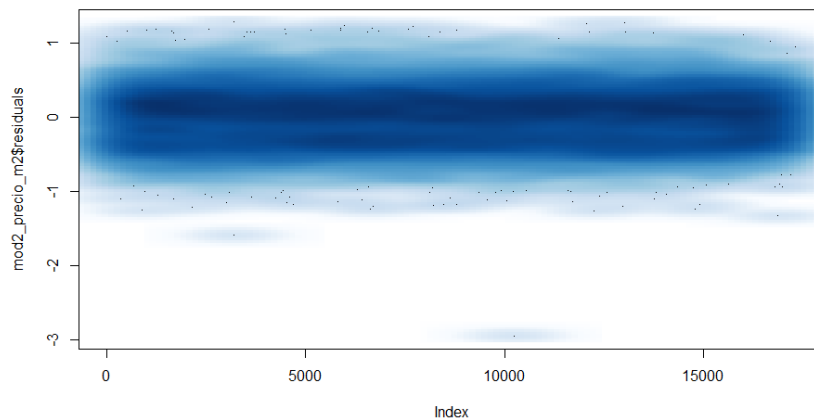


```
plot(mod2_precio_m2$residuals)
```

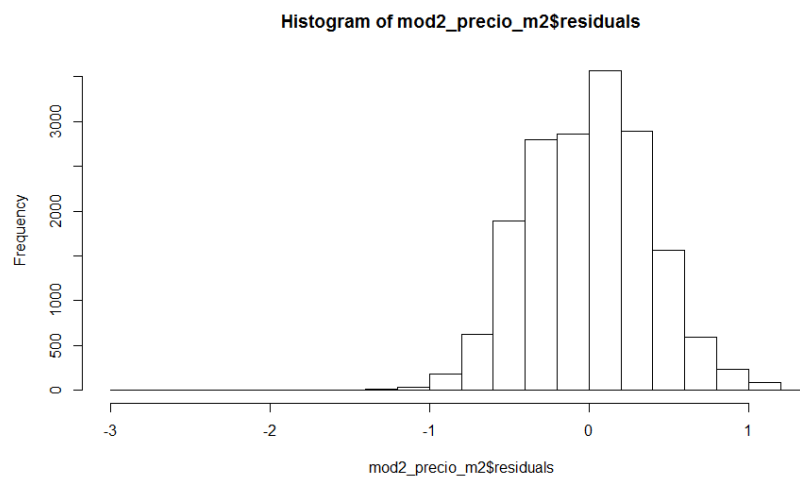




`smoothScatter(mod2_preco_m2$residuals)`



`hist(mod2_preco_m2$residuals)`



Y el interval de confianza:

`confint(mod2_preco_m2,level=0.95)`

```
> confint(mod2_preco_m2,level=0.95)
                2.5 %      97.5 %
(Intercept) 1.220667e+01 1.223453e+01
sqft_living  3.907338e-04 4.029721e-04
> |
```

Una pequeña comprobación:

Modelo	Variable Dependiente	Variable Independiente	Interpretación
Regresión Log-Level $\ln(y) = \beta_0 + \beta_1 x + \epsilon$	$\ln(y)$	x	Un aumento de 1 unidad en x se corresponde con un aumento del 100*beta% en y. (semielasticidad)

`log(1.2220+(0.0003970*2)) - log(1.2220+(0.0003970*1))`

```
> log(1.2220+(0.0003970*2)) - log(1.2220+(0.0003970*1))
[1] 0.000324719
```



Vamos a comparar los dos modelos:

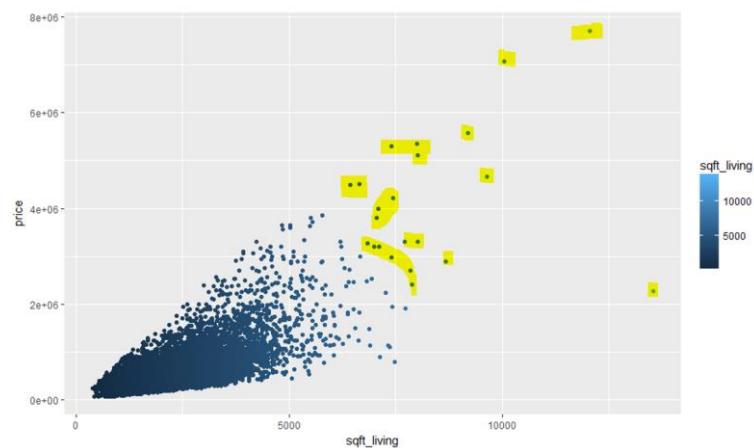
```
AIC(mod1Precio_m2)
```

```
AIC(mod2Precio_m2)
```

```
> AIC(mod1Precio_m2)
[1] 482205.1
> AIC(mod2Precio_m2)
[1] 15561.29
> |
```

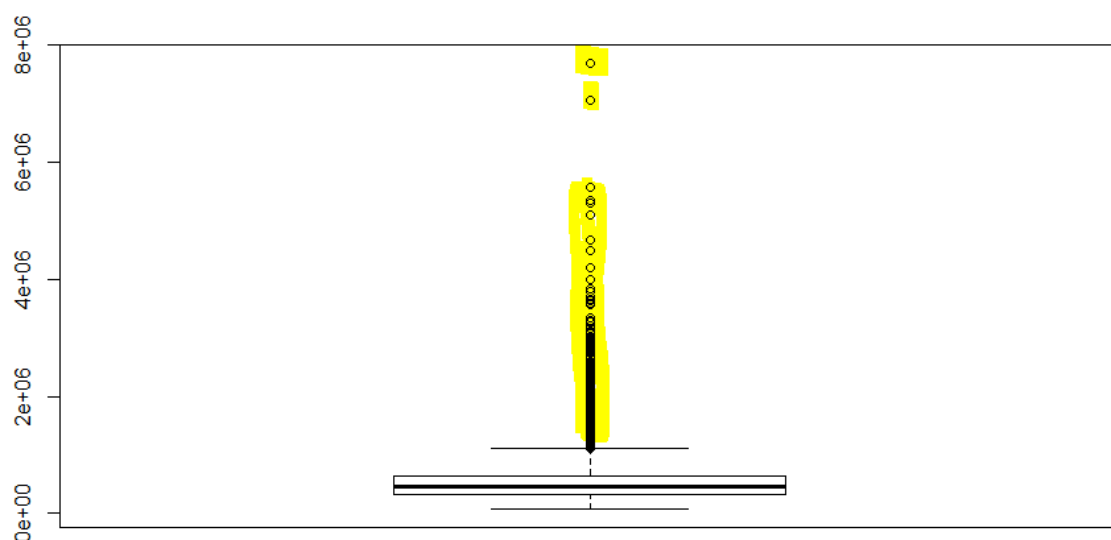
Vemos que el segundo modelo, tal y como comprobamos con la distribución de los residuos, es bastante mejor.

No obstante en los gráficos se ve unos outliers muy pronunciados:



Lo comprobamos con un boxplot:

```
boxplot(Precio_m2$price)
```



Pues va a ser que si, por lo que vamos a hacer un modelo con estadística robusta:



5.3 Modelo con estadística robusta

Creamos el nuevo modelo utilizando estadística robusta:

```
if (!require("MASS")){
  install.packages("MASS")
  library(MASS)
}

if (!require("caTools")){
  install.packages("caTools")
  library(caTools)
}
```

```
mod3_precio_m2=rlm(price~sqft_living,data=precio_m2)
```

```
summary(mod3_precio_m2)
```

```
Call: rlm(formula = price ~ sqft_living, data = precio_m2)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-877308 -122597   -6608   117687  5033741
```

```
Coefficients:
```

```
              Value      Std. Error t value
(Intercept) 58629.8045   3438.8487   17.0493
sqft_living   216.4008     1.5109   143.2246
```

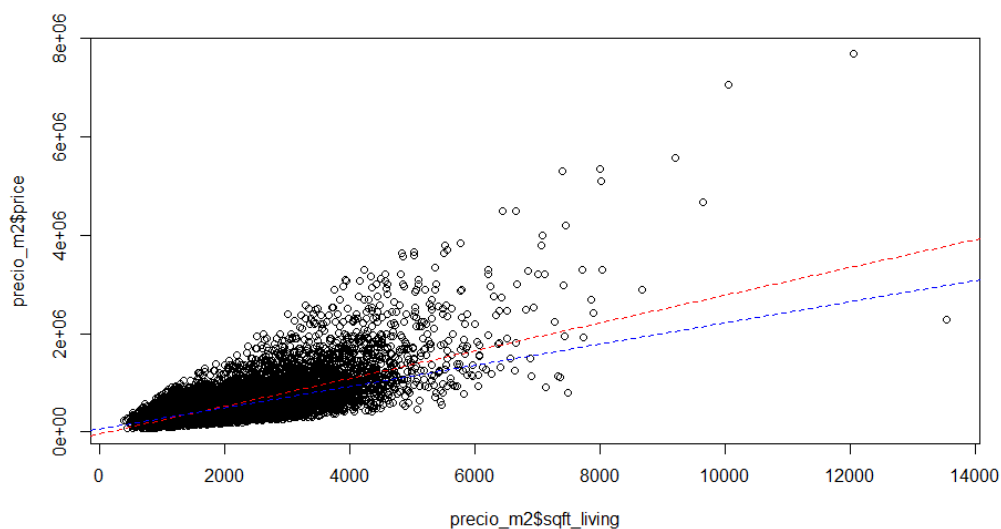
```
Residual standard error: 178500 on 17344 degrees of freedom
```

Vamos a pintar el 1er (estadística normal) y 3r modelo (estadística robusta):

```
plot(precio_m2$sqft_living,precio_m2$price)
```

```
abline(mod1_precio_m2,col="red",lty = "dashed")
```

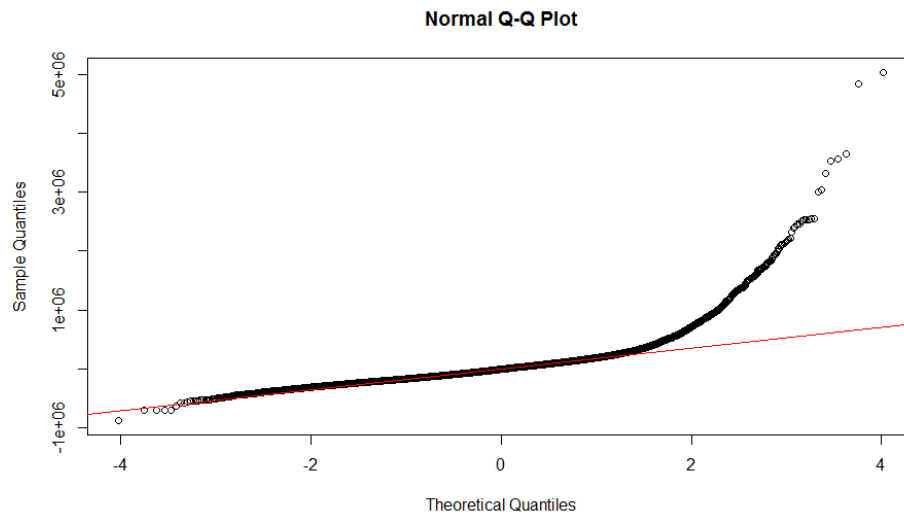
```
abline(mod3_precio_m2,col="blue",lty = "dashed")
```



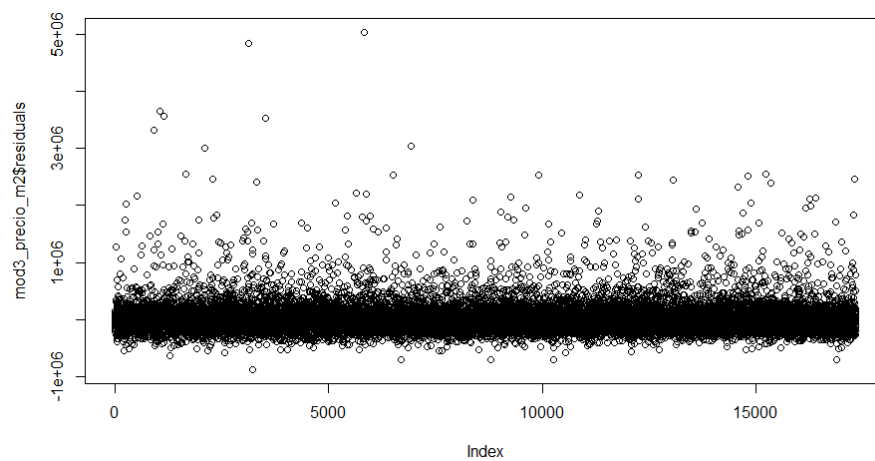


Ya vemos que se ha corregido la linea de regresión. Vamos a ver los residuos:

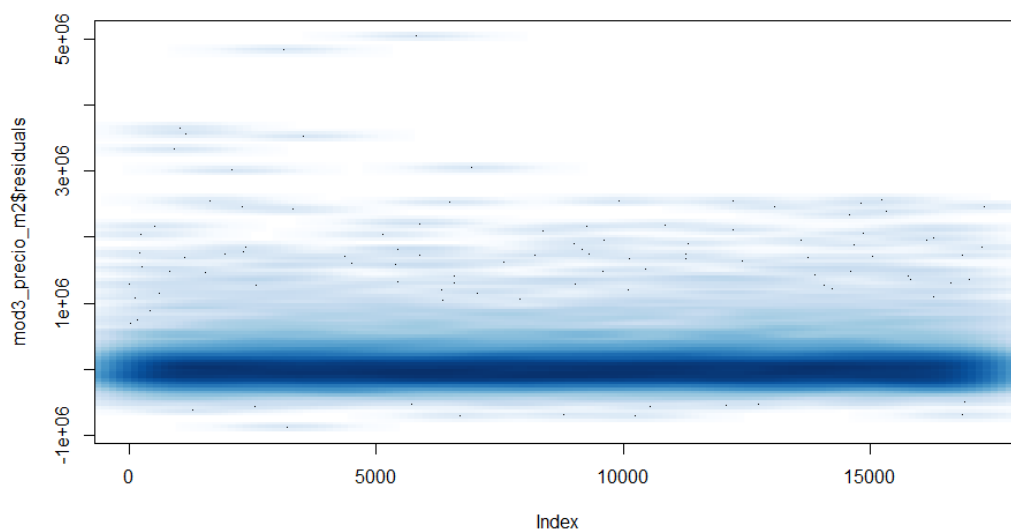
```
qqnorm(mod3_precio_m2$residuals); qqline(mod3_precio_m2$residuals,col=2)
```



```
plot(mod3_precio_m2$residuals)
```

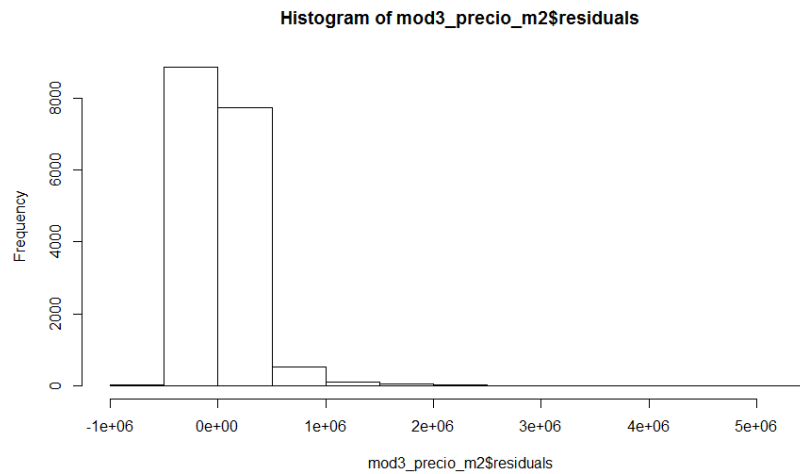


```
smoothScatter(mod3_precio_m2$residuals)
```





```
hist(mod2_precio_m3$residuals)
```



```
confint.default(mod3_precio_m2,level=0.95)
```

```
> confint.default(mod3_precio_m2,level=0.95)
              2.5 %      97.5 %
(Intercept) 51889.7849 65369.8242
sqft_living   213.4394   219.3621
> |
```

5.4 Modelo de porcentaje con estadística robusta

```
mod4_precio_m2=rlm(log(price)~sqft_living, data=precio_m2)
```

```
summary(mod4_precio_m2)
```

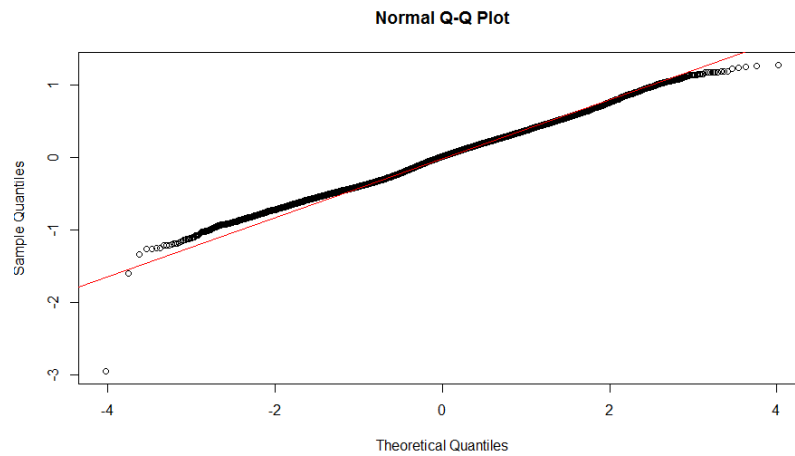
```
Call: rlm(formula = log(price) ~ sqft_living, data = precio_m2)
Residuals:
    Min       1Q   Median       3Q      Max
-2.94692 -0.28597  0.01777  0.26397  1.28276

Coefficients:
              value      Std. Error t value
(Intercept)  12.2189      0.0073  1680.7614
sqft_living   0.0004      0.0000  124.1117

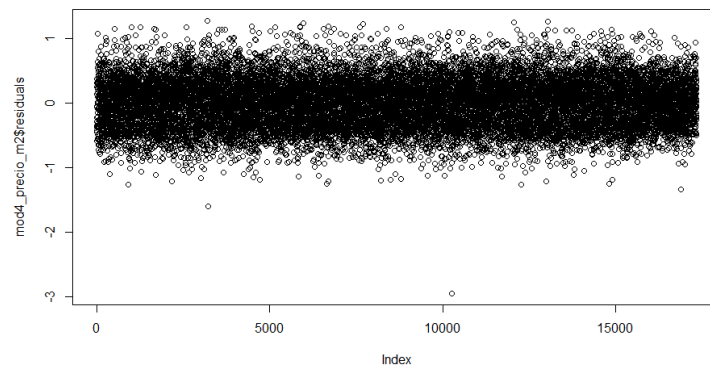
Residual standard error: 0.4058 on 17344 degrees of freedom
```



```
qqnorm(mod4_precio_m2$residuals); qqline(mod4_precio_m2$residuals,col=2)
```

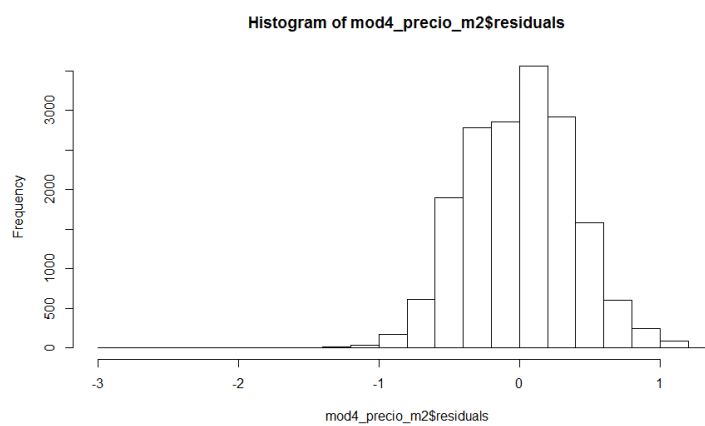


```
plot(mod4_precio_m2$residuals)
```



```
smoothScatter(mod4_precio_m2$residuals)
```

```
hist(mod4_precio_m2$residuals)
```



```
confint.default(mod4_precio_m2,level=0.95)
```

```
> confint.default(mod4_precio_m2,level=0.95)
              2.5 %      97.5 %
(Intercept) 1.220468e+01 1.223317e+01
sqft_living  3.901709e-04 4.026918e-04
> |
```



6. Comparación modelos a datos de test

Comparamos los cuatro modelos

`AIC(mod1_precio_m2)`

`AIC(mod2_precio_m2)`

`AIC(mod3_precio_m2)`

`AIC(mod4_precio_m2)`

```
> AIC (mod4_precio_m2)
[1] 15562.1
> AIC(mod1_precio_m2)
[1] 482205.1
> AIC(mod2_precio_m2)
[1] 15561.29
> AIC(mod3_precio_m2)
[1] 483321.2
> AIC(mod4_precio_m2)
[1] 15562.1
```



Curiosamente el segundo modelo, no los de estadística robusta, es el mejor.



7. Modelo predictivo

Empezamos analizando las variables que tenemos:

`cor(house_train)`

```
id          yr_renovated  zipcode      lat      long  sqft_living15  sqft_lot15  Year
price      -0.0130987868 -0.009337180 -0.0067382920 0.021230716 -0.007654411 -0.138246129 0.007386246
bedrooms   0.1234095723 -0.054531195 0.3092374255 0.021163913 0.583278227 0.081183829 0.005317094
bathrooms  0.0137042374 -0.159859242 -0.0111918712 0.137447190 0.403854695 0.029268349 -0.005434434
sqft_living 0.0462169277 -0.204065182 0.0261098248 0.222031720 0.568176584 0.085534210 -0.026383538
sqft_lot   0.0536189448 -0.196257644 0.0564593882 0.237417585 0.756152663 0.179463867 -0.028033012
sqft_lot   0.0070903984 -0.127928980 -0.0853239793 0.226234877 0.147866125 0.727361467 0.001705549
floors     -0.0007488357 -0.057711135 0.0494071532 0.125551101 0.280486222 -0.006984166 -0.016851329
waterfront 0.0968730997 0.027847388 -0.0138659592 -0.038825257 0.085468987 0.024038149 -0.006916734
view       0.1030421334 -0.085186570 0.0087692939 -0.076203004 0.279626531 0.069096549 0.001402751
condition -0.0611575958 0.007247845 -0.0136670402 -0.106005009 -0.094565233 -0.002249405 -0.046827724
grade      0.0059913890 -0.184548647 0.1140598078 0.195897271 0.711139989 0.121607543 -0.028022728
sqft_above 0.0216182855 -0.258513252 0.0009652095 0.341327310 0.732695609 0.189630762 -0.021983436
sqft_basement 0.0709489126 0.074254076 0.1152870628 -0.143064203 0.204103724 0.019150946 -0.017209760
yr_built   -0.2222674130 -0.347849962 -0.1508830212 0.408501032 0.326799638 0.072187656 0.004433015
yr_renovated 1.0000000000 0.065068244 0.0297252360 -0.069851926 -0.003437213 0.008460143 -0.029370210
zipcode    0.0650682439 1.0000000000 0.2644200181 -0.560997922 -0.275857224 -0.145852754 -0.001597465
lat        0.0297252360 0.264420018 1.0000000000 -0.133764788 0.049638321 -0.089660266 -0.029409322
long       -0.0698519258 -0.560997922 -0.133764788 1.000000000 0.334420241 0.255399710 -0.002011954
sqft_living15 0.0034372128 -0.275857224 0.0496383213 0.334420241 1.000000000 0.184974934 -0.018898012
sqft_lot15 0.0084601426 -0.145852754 -0.0896602659 0.255399710 0.184974934 1.000000000 0.003300556
year       -0.0293702095 -0.001597465 -0.0294093223 -0.002011954 -0.018898012 -0.003300556 1.000000000
```

../..

Vemos que la suma entre `sqft_above`, `sqft_basement` nos da `sqft_living` por lo que descartaremos las dos primeras

Vemos Multicolinealidad entre `bedroom` y `batchroom` y `sqft_living`. Quitamos las dos primeras.

Latitud y longitud quitamos porque vamos a utilizar el `zipcode`, que creo los comprende.

Quitamos el año de renovación porque hay muy pocas lecturas.

`sqft_lot15`, `sqft_living15` y `grade` desconocemos pues quitamos.

Finalmente `year` no lo utilizaré. Soy lo peor, limpio ese dato pá no usarlo.... 🤔 Es para que el señor de la vara de los data scientist me dé bien ...

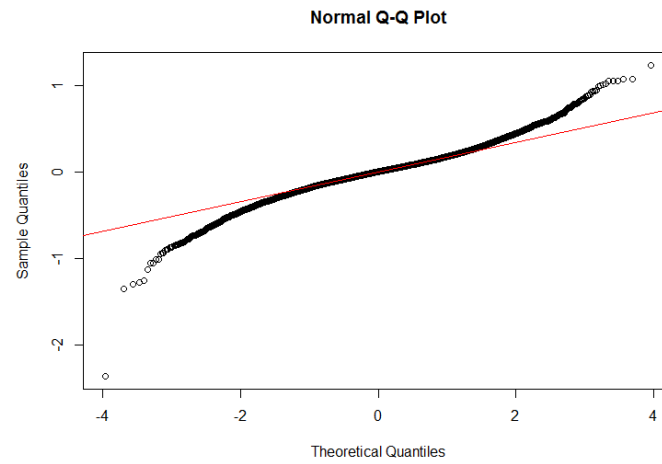
Dejamos el resto.

```
mod5_predict=lm(log(price)~sqft_living+sqft_lot+waterfront+view+condition+
+zipcode+yr_renovated, data=train)
```

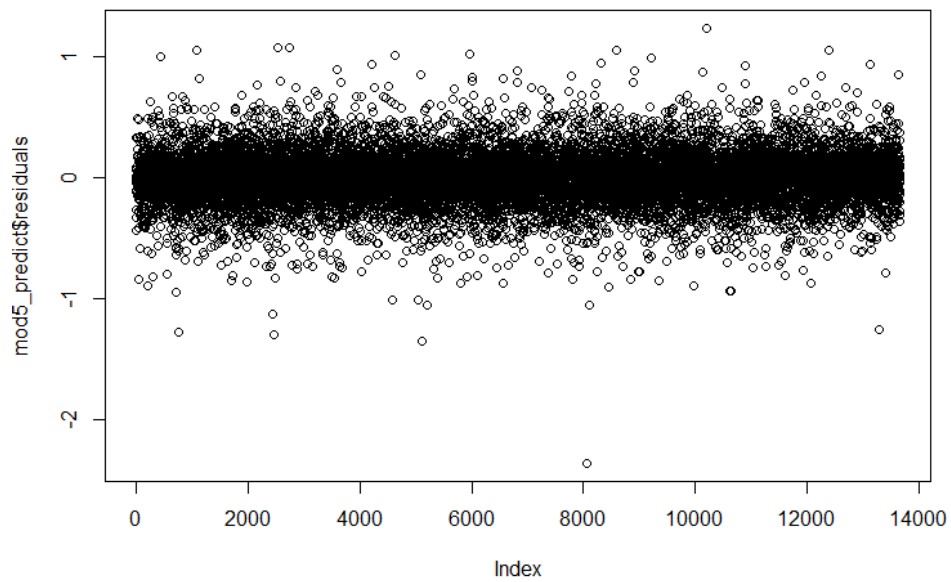
```
summary(mod5_predict)
```

```
Residual standard error: 0.2122 on 17270 degrees of freedom
Multiple R-squared: 0.8383, Adjusted R-squared: 0.8376
F-statistic: 1194 on 75 and 17270 DF, p-value: < 2.2e-16
```

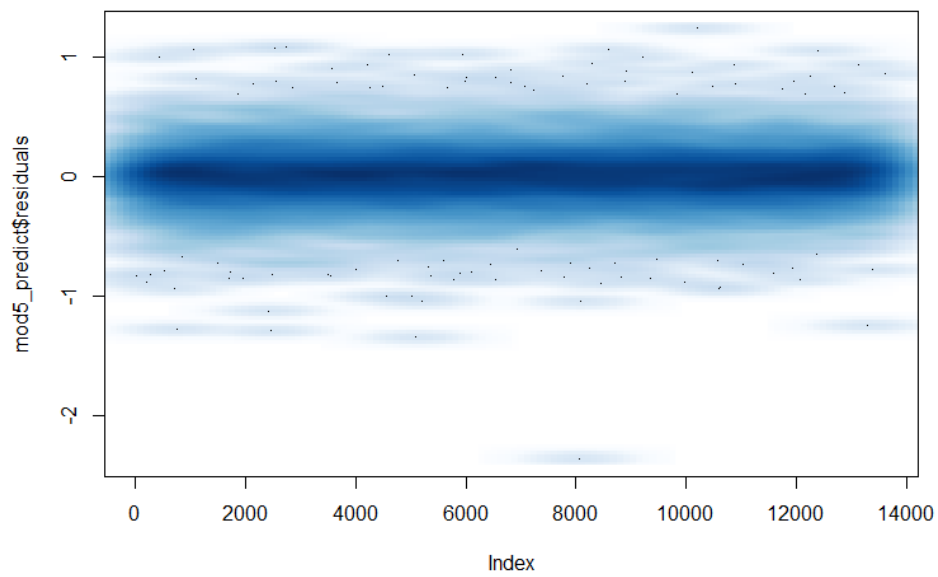
```
qqnorm(mod5_predict$residuals); qqline(mod5_predict$residuals,col=2)
```

`plot(mod5_predict$residuals)`

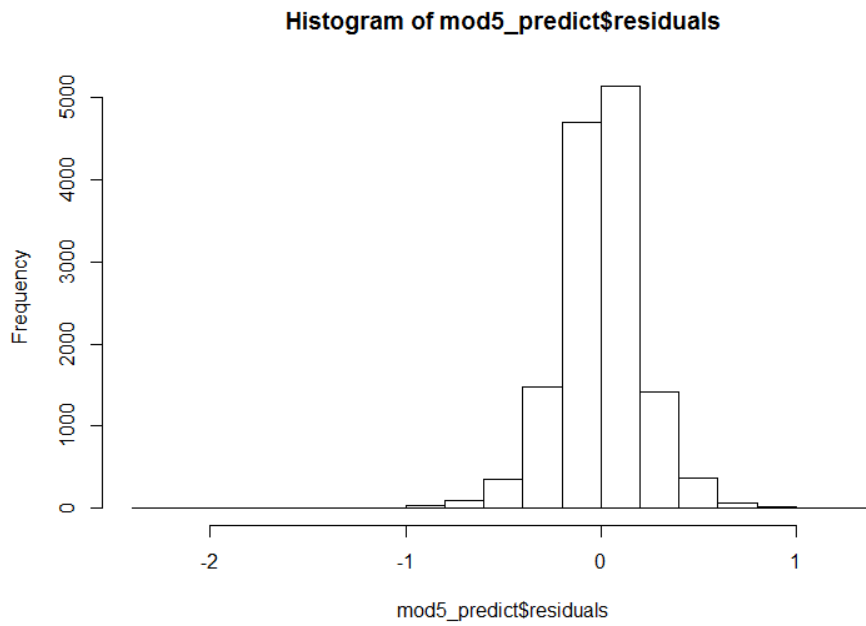


`smoothScatter(mod5_predict$residuals)`





```
hist(mod5_predict$residuals)
```



Comprobamos el modelo

```
AIC(mod5_predict)
```

```
> AIC(mod5_predict)
[1] -4479.177
```



8. Aplicación modelos a datos de test

Vamos a aplicar el modelo predictivo que hemos hecho al CSV que tenemos con casas sin precio.

```
house_test=read.csv("house_test.csv")
```

```
house_test$waterfront=as.factor(house_test$waterfront)
```

```
house_test$condition=as.factor(house_test$condition)
```

```
house_test$zipcode=as.factor(house_test$zipcode)
```

```
house_test$price=exp(predict(mod5_predict,newdata=house_test,type="response"))
```

```
summary(house_test)
```

```
write.csv(house_test, "house_test_valorado.csv")
```

9. Entregables

8.1 Documento de construcción del modelo

Es el presente documento word

8.2 Documento de análisis del efecto de la superficie de la vivienda en el precio

Documento R



Entregable Actividad
1.R

8.3 Fichero Test con columna de precio resultante

Fichero house_test.csv



house_test_valorado.
csv

