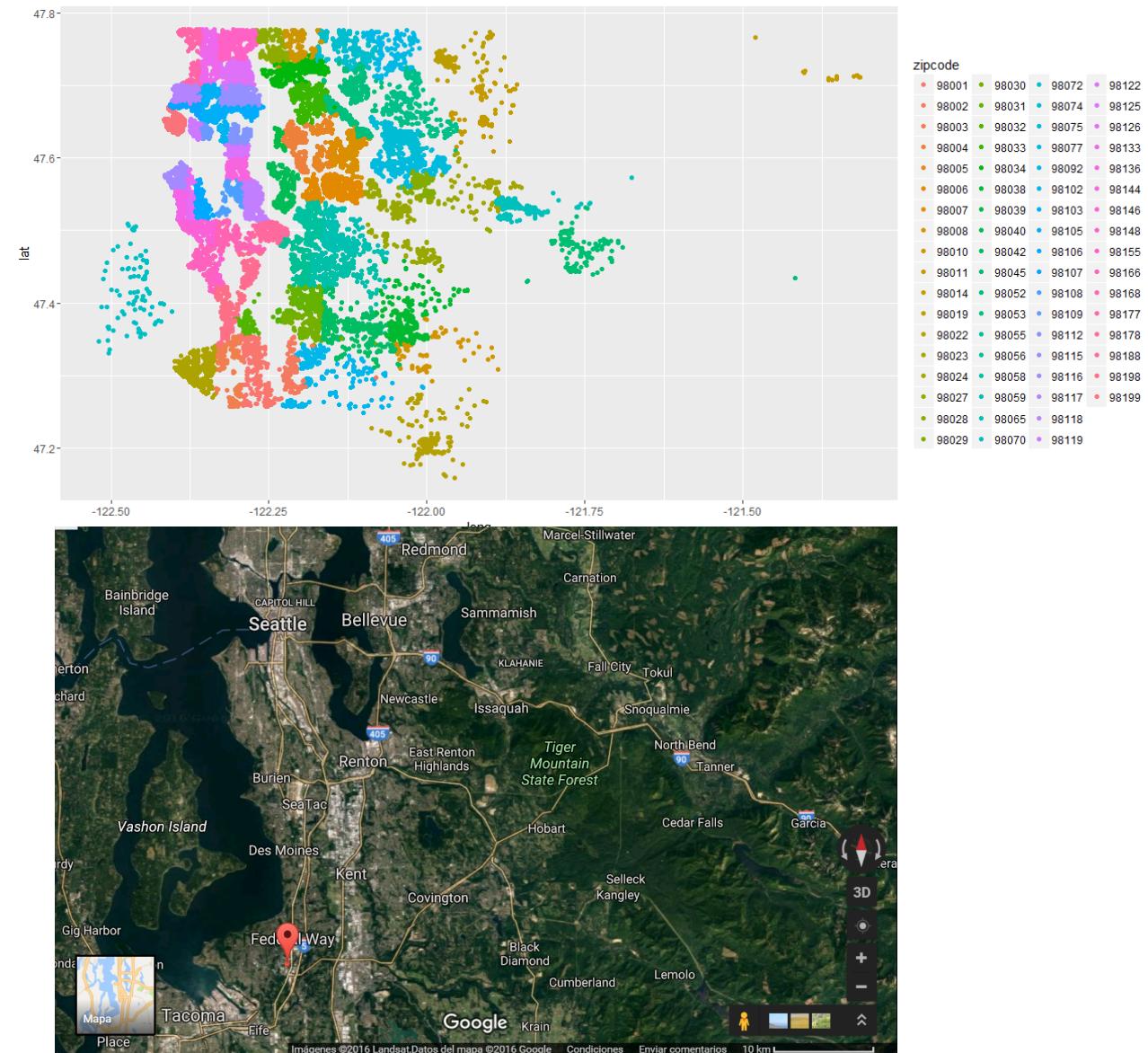


Primer Ejercicio: Construcción de un modelo de regresión.

Análisis previo

En primer lugar analizaré los datos de las viviendas del data set.

Graficando la ubicación y el código postal de las viviendas podemos observar que todas se encuentran en Seattle.



Voy a analizar otras características de las viviendas, empezando por los campos conocidos. Adicionalmente averiguaré el significado de las demás variables para poder evaluar su utilidad en el modelo.

Las variables conocidas son:

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
yr_built: año de construcción
yr_renovated: año de reforma
zipcode: código postal
lat: latitud
long: longitud

Busco el significado de las variables desconocidas:

condition:

Relative to age and grade. Coded 1-5.

1 = Poor- Worn out. Repair and overhaul needed on painted surfaces, roofing, plumbing, heating and numerous functional inadequacies. Excessive deferred maintenance and abuse, limited value-in-use, approaching abandonment or major reconstruction; reuse or change in occupancy is imminent. Effective age is near the end of the scale regardless of the actual chronological age.

2 = Fair- Badly worn. Much repair needed. Many items need refinishing or overhauling, deferred maintenance obvious, inadequate building utility and systems all shortening the life expectancy and increasing the effective age.

3 = Average- Some evidence of deferred maintenance and normal obsolescence with age in that a few minor repairs are needed, along with some refinishing. All major components still functional and contributing toward an extended life expectancy. Effective age and utility is standard for like properties of its class and usage.

4 = Good- No obvious maintenance required but neither is everything new. Appearance and utility are above the standard and the overall effective age will be lower than the typical property.

5= Very Good- All items well maintained, many having been overhauled and repaired as they have shown signs of wear, increasing the life expectancy and lowering the effective age with little deterioration or obsolescence evident with a high degree of utility.

grade:

Represents the construction quality of improvements. Grades run from grade 1 to 13. Generally defined as:

1-3 Falls short of minimum building standards. Normally cabin or inferior structure.

4 Generally older, low quality construction. Does not meet code.

5 Low construction costs and workmanship. Small, simple design.

6 Lowest grade currently meeting building code. Low quality materials and simple designs.

7 Average grade of construction and design. Commonly seen in plats and older sub-divisions.

8 Just above average in construction and design. Usually better materials in both the exterior and interior finish work.

9 Better architectural design with extra interior and exterior design and quality.

10 Homes of this quality generally have high quality features. Finish work is better and more design quality is seen in the floor plans. Generally have a larger square footage.

11 Custom design and higher quality finish work with added amenities of solid woods, bathroom fixtures and more luxurious options.

12 Custom design and excellent builders. All materials are of the highest quality and all conveniences are present.

13 Generally custom designed and built. Mansion level. Large amount of highest quality cabinet work, wood trim, marble, entry ways etc.

sqft_above: = sqft_living - sqft_basement

sqft_basement:

sqft_living15: the average house square footage of the 15 closest neighbours

sqft_lot15: the average lot square footage of the 15 closest neighbours

Fuente:

<http://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r>

<https://www.kaggle.com/harlfoxem/d/harlfoxem/housesalesprediction/>

Voy a formatear el data set, para que las variables categóricas se representen como factores.

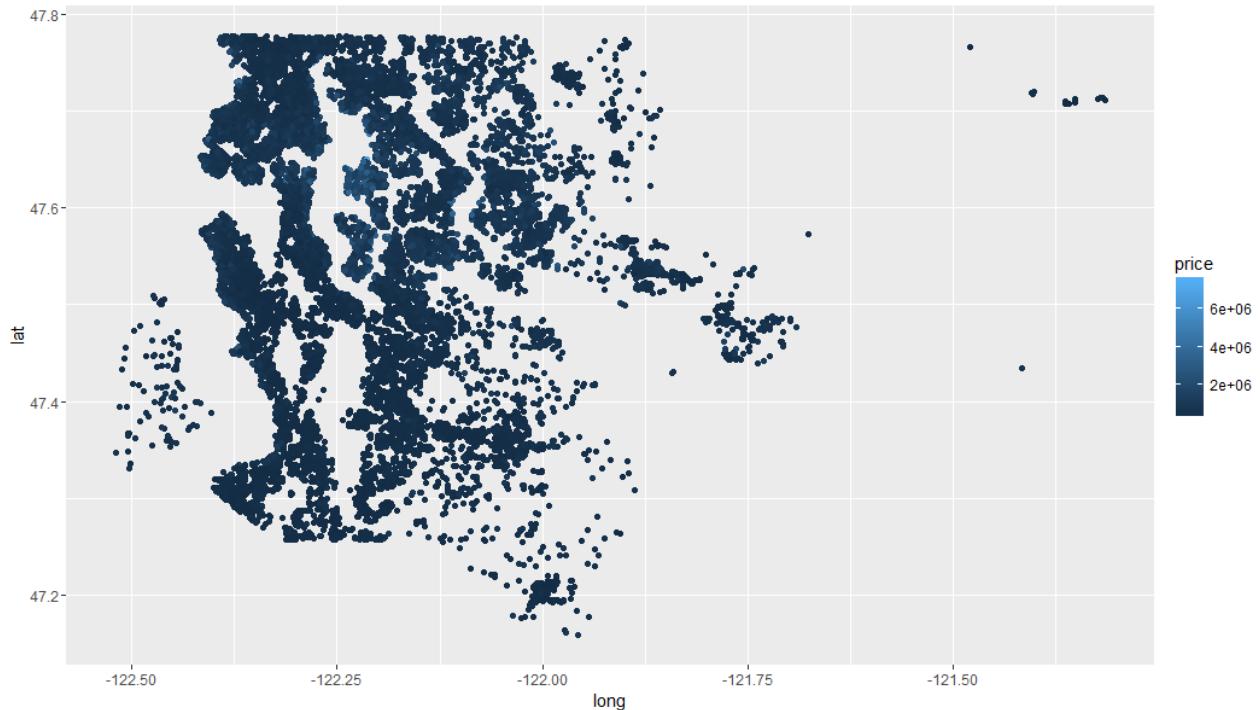
Tomo las siguientes variables como categóricas: bedrooms, bathrooms, floors, waterfront, view, condition, grade

El resto de las variables interesantes las dejo en formato numérico:

yr_built, yr_renovated, price, sqft_living , sqft_lot.

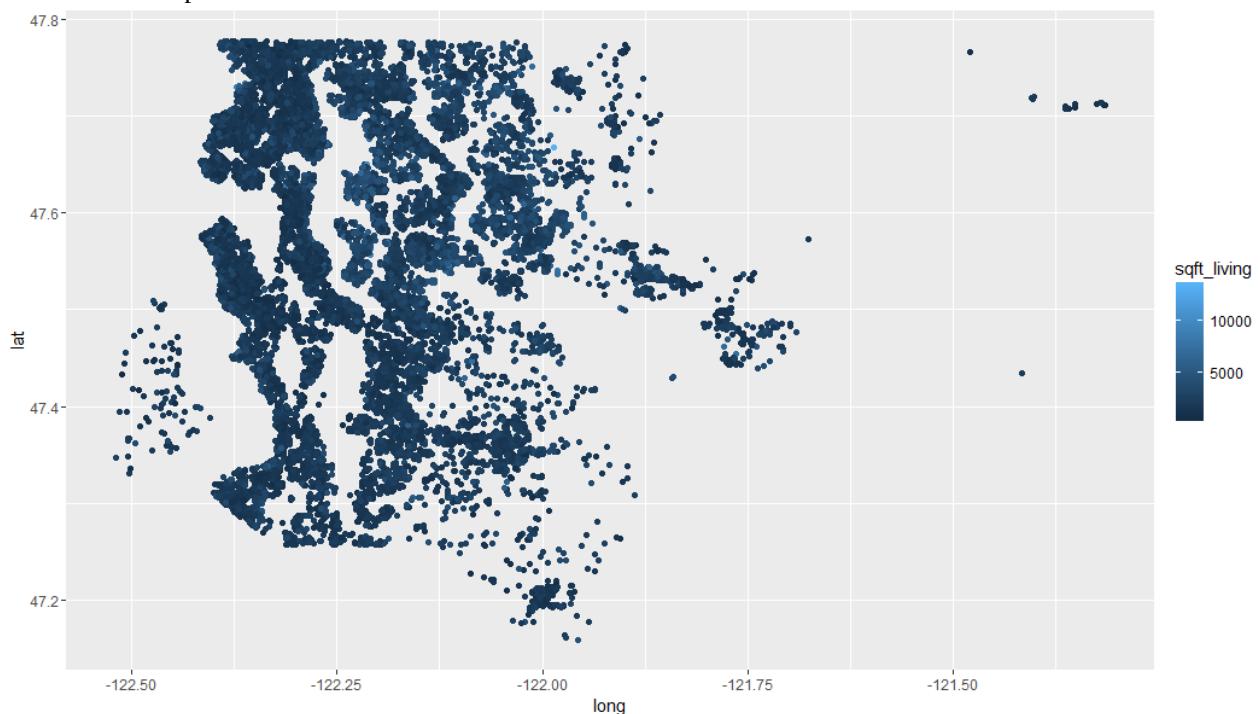
Análisis del efecto de la superficie de la vivienda en el precio de la vivienda

Veamos los precios de las viviendas:



El precio no parece tener mucha dispersión por zona, excepto el centro en alrededores de Capitol Hill y barrio Bellevue. Ahí las casas son mucho mas caras.

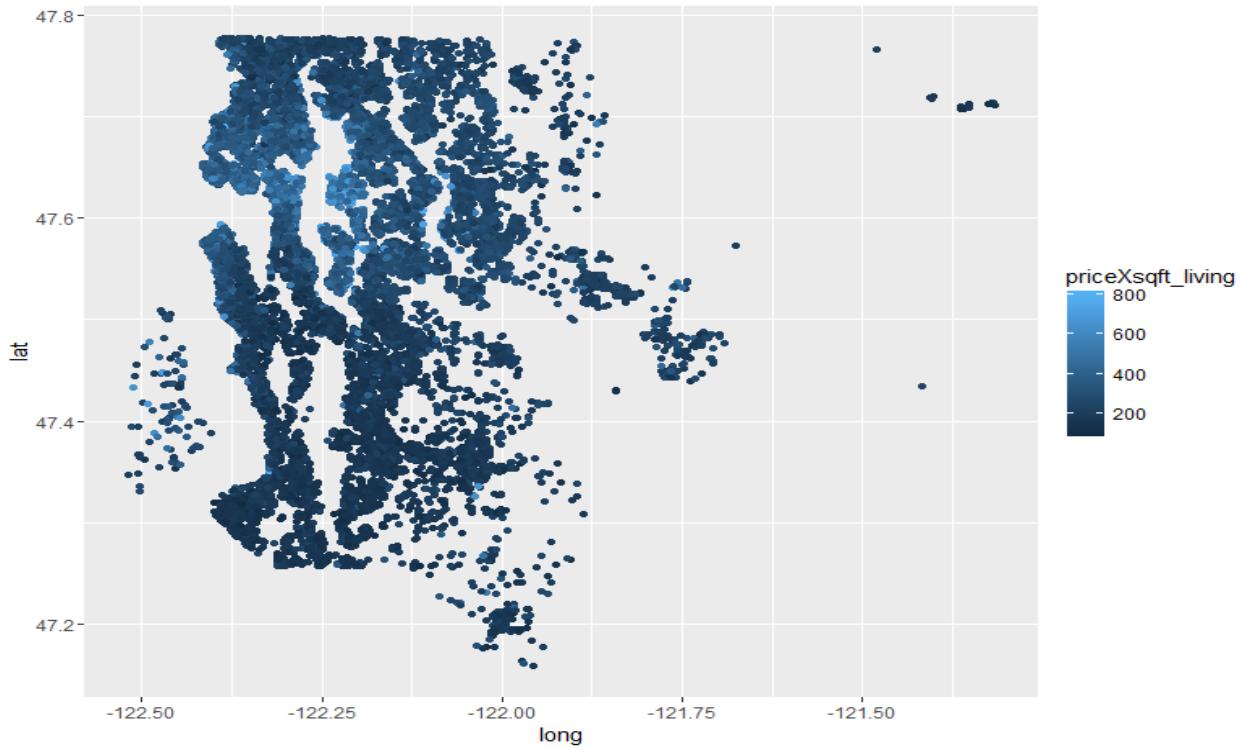
Analizando la superficie:



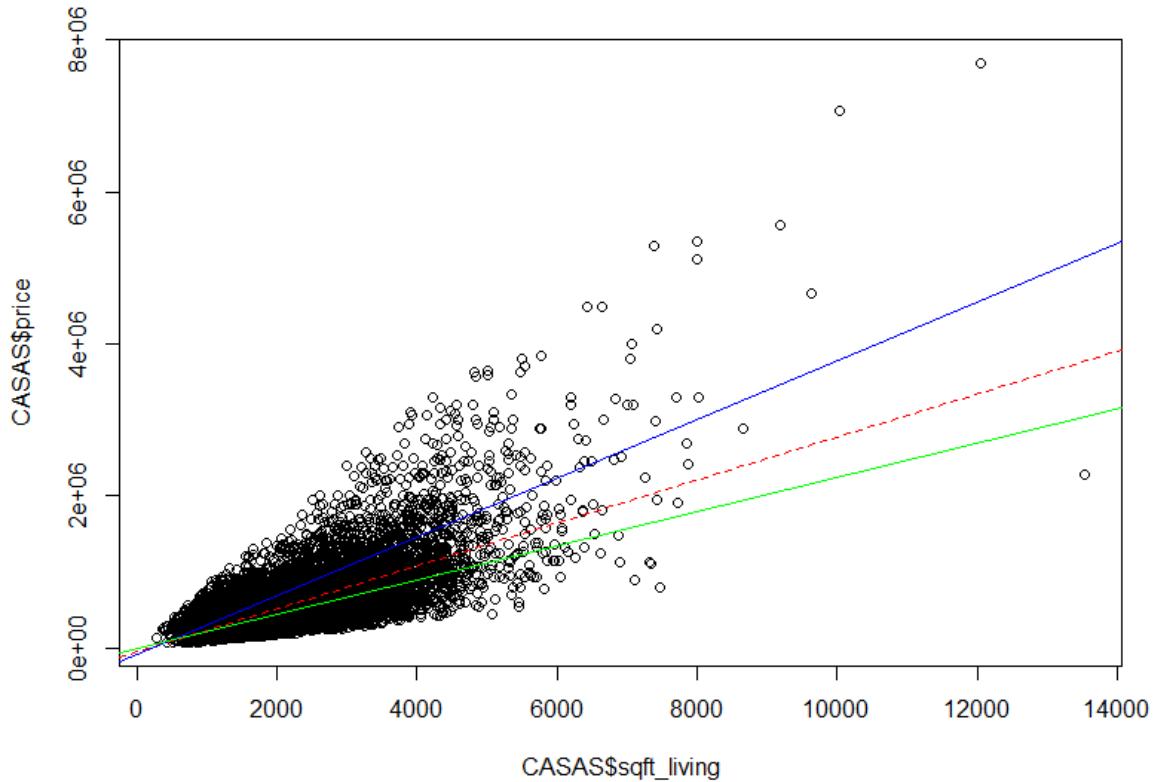
Se puede ver que en la zona cara las casas son más grandes. En general las casas son más grandes en el este y más pequeñas en el oeste.

Podemos observar que el precio no tiene relación directamente con la superficie. En algunas zonas las casas son pequeñas pero caras. Con toda seguridad la superficie no es el único factor del que depende el precio de las viviendas.

Veamos el precio por metro cuadrado (más bien por pie al cuadrado):



Es fácil de ver que hay una zona más cara que las demás. En alrededores del centro cada metro de superficie cuesta mucho más que en las demás zonas. Esta métrica indica bastante sobre la relación sobre el precio y la superficie marginal que queremos obtener. Es posible que existan dos patrones diferentes. Voy a comprobarlo:



El gráfico representa la relación entre el precio y superficie de la vivienda:

en rojo – para todas las viviendas

en azul – viviendas en el centro

en verde – viviendas en afuera

Podemos observar que la pendiente que determina la relación entre la superficie y el precio cambia dependiendo de la ubicación de la vivienda. En el caso de las viviendas del centro cada metro cuadrado adicional es más caro que en el caso de las viviendas que están en afuera.

Adicionalmente realizaré un chow test para confirmar el cambio de la estructura:

```
chow.test(CASAS_CENTRO$price, CASAS_CENTRO$sqft_living, CASAS_AFUERAS$price, CASAS_AFUERAS$sqft_living)
```

F value	d.f.1	d.f.2	P value
4051.488	2.000	17380.000	0.000

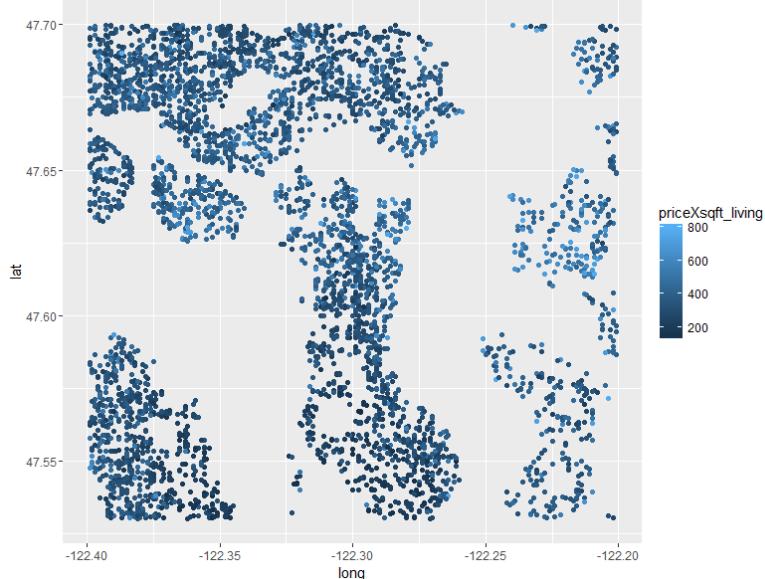
```
chow.test(log(CASAS_CENTRO$price), CASAS_CENTRO$sqft_living, log(CASAS_AFUERAS$price), CASAS_AFUERAS$sqft_living)
```

F value	d.f.1	d.f.2	P value
3828.976	2.000	17380.000	0.000

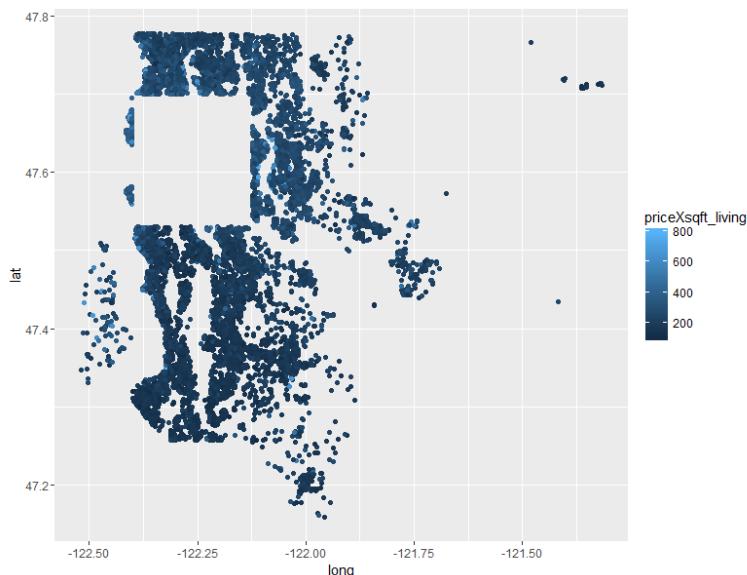
Se confirma la diferencia estructural y concluyo que los dos data sets probablemente tengan elasticidad del precio a la variación de superficie diferente.

Voy a dividir la población en dos partes:

Viviendas en el centro:



Viviendas en afuera:



Construcción del modelo básico para las viviendas del centro y para las viviendas en afuera.

Como el objetivo del análisis es construir un modelo que refleje el efecto de superficie en el precio de la vivienda voy a construir un modelo log-level. De esta forma podré estimar el impacto marginal de cada pie adicional de superficie a la variación del precio.

Para las casas del centro construyo el siguiente modelo básico:

```
modelocentrolog=lm(log(price)~sqft_living,data=CASAS_CENTRO)
```

```
lm(formula = log(price) ~ sqft_living, data = CASAS_CENTRO)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.44613	-0.16977	0.00606	0.18001	0.96154

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.253e+01	8.915e-03	1405.2	<2e-16 ***
sqft_living	3.963e-04	3.928e-06	100.9	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2929 on 5700 degrees of freedom

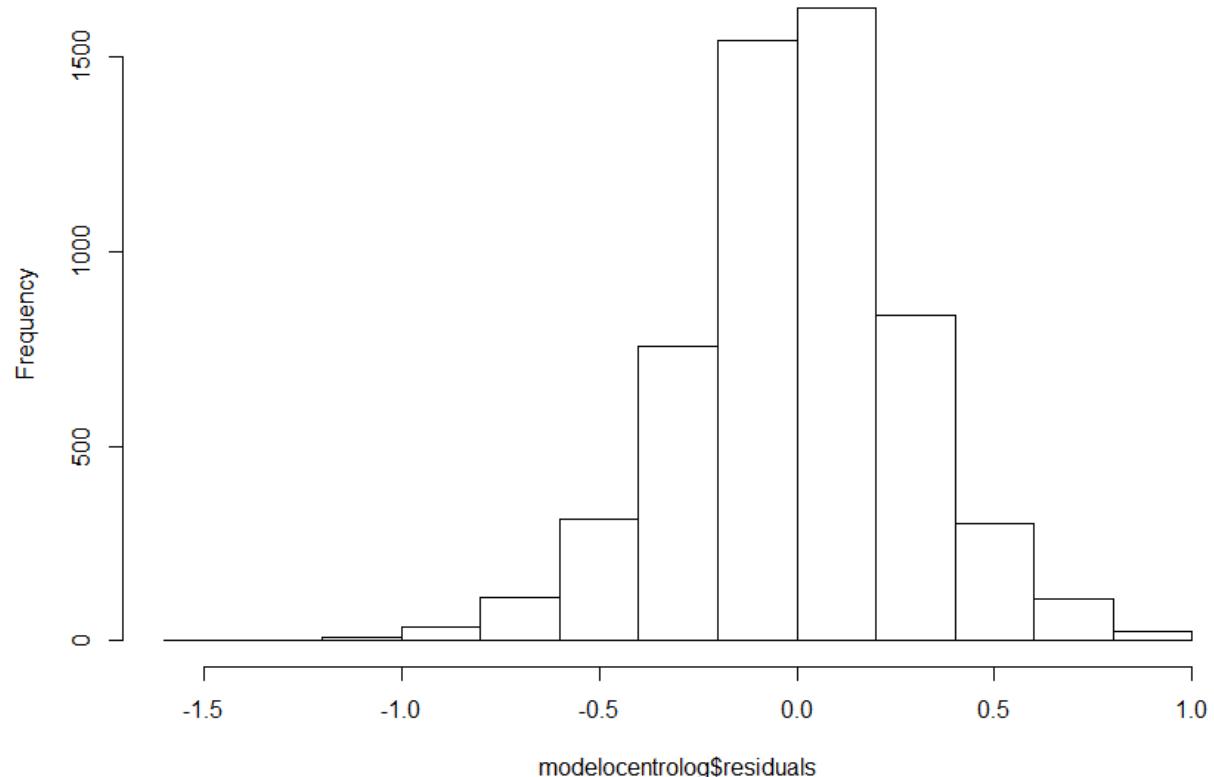
Intervalo de confianza:

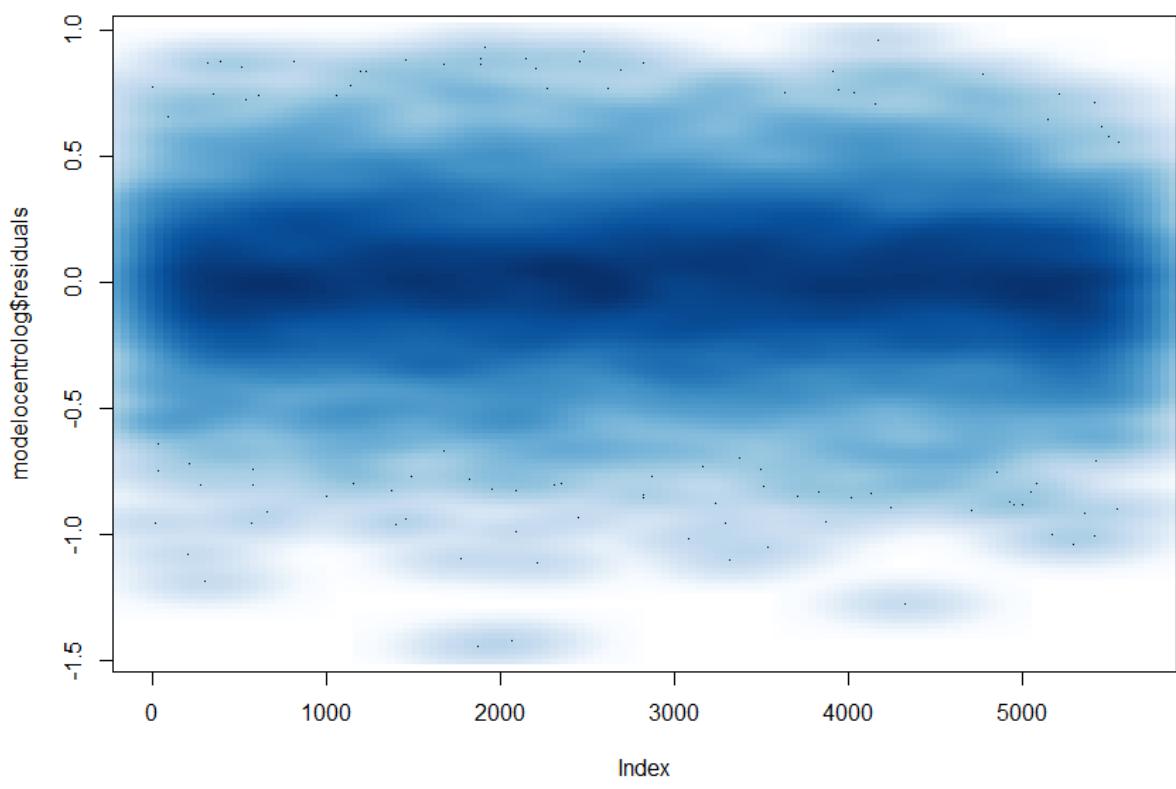
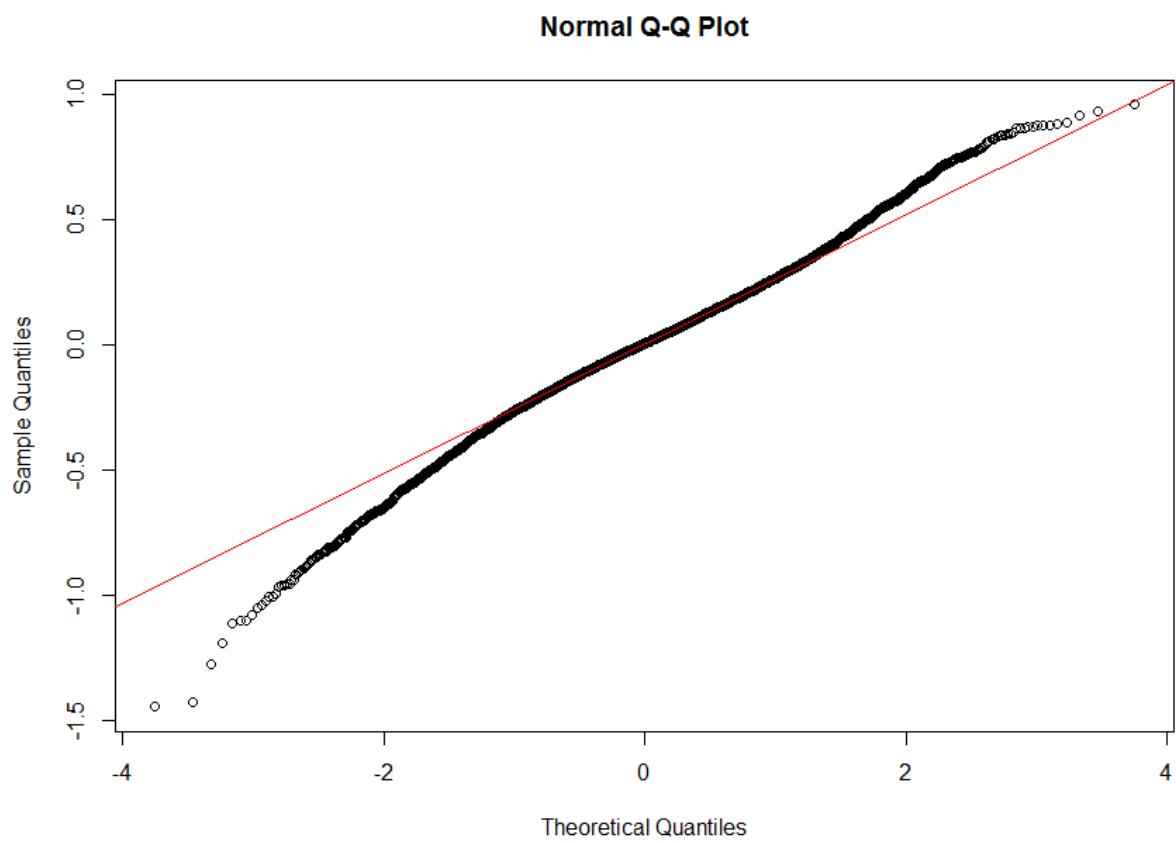
	2.5 %	97.5 %
(Intercept)	1.250951e+01	1.254447e+01
sqft_living	3.886385e-04	4.040379e-04

Esto significa que bajo nivel de confianza de 95% un ascenso de la superficie de un piso de 1000 pies cuadrados por 1 pie está asociado con un ascenso de precio entre 155.3191 y 169.8147 dólares.

Residuos tienen media cercana a cero y distribución aproximada a normal en el segundo y tercer quartil teórico, aunque observo una desviación en las colas:

Histogram of modelocentrolog\$residuals





Voy a comprobar si el desajuste del modelo podría ser por outliers, construyendo un modelo robusto para comparar.

```
Call: rlm(formula = log(price) ~ sqft_living, data = CASAS_CENTRO)
Residuals:
```

Min	1Q	Median	3Q	Max
-1.467173	-0.173288	0.002145	0.176352	0.956576

Coefficients:

	value	Std. Error	t value
(Intercept)	12.5277	0.0085	1475.2372
sqft_living	0.0004	0.0000	106.383

Residual standard error: 0.2589 on 5663 degrees of freedom

Intervalo de confianza:

	2.5 %	97.5 %
(Intercept)	1.251108e+01	12.544370007
sqft_living	3.906901e-04	0.000405356

El error del modelo robusto baja pero no es una bajada considerable. El intervalo de confianza también es mas estrecho, pero no mucho más.

Comparando la bondad de ambos modelos obtengo:

	modelocentrolog	modelocentrologrobusto
AIC	2181.165	2182.511
BIC	2201.111	2202.456

Concluyo que el modelo no mejora mediante tratamiento de outliers. Más adelante voy a ver si se puede mejorar añadiendo más variables.

Para las casas de las afueras construyo el siguiente modelo básico:

```
modeloafueras=lm(log(price)~sqft_living,data=CASAS_AFUERAS)
```

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

```
Residuals:
```

Min	1Q	Median	3Q	Max
-2.93030	-0.23043	-0.00428	0.22854	1.41924

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.205e+01	7.653e-03	1574.4	<2e-16 ***
sqft_living	4.078e-04	3.360e-06	121.4	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3222 on 11680 degrees of freedom

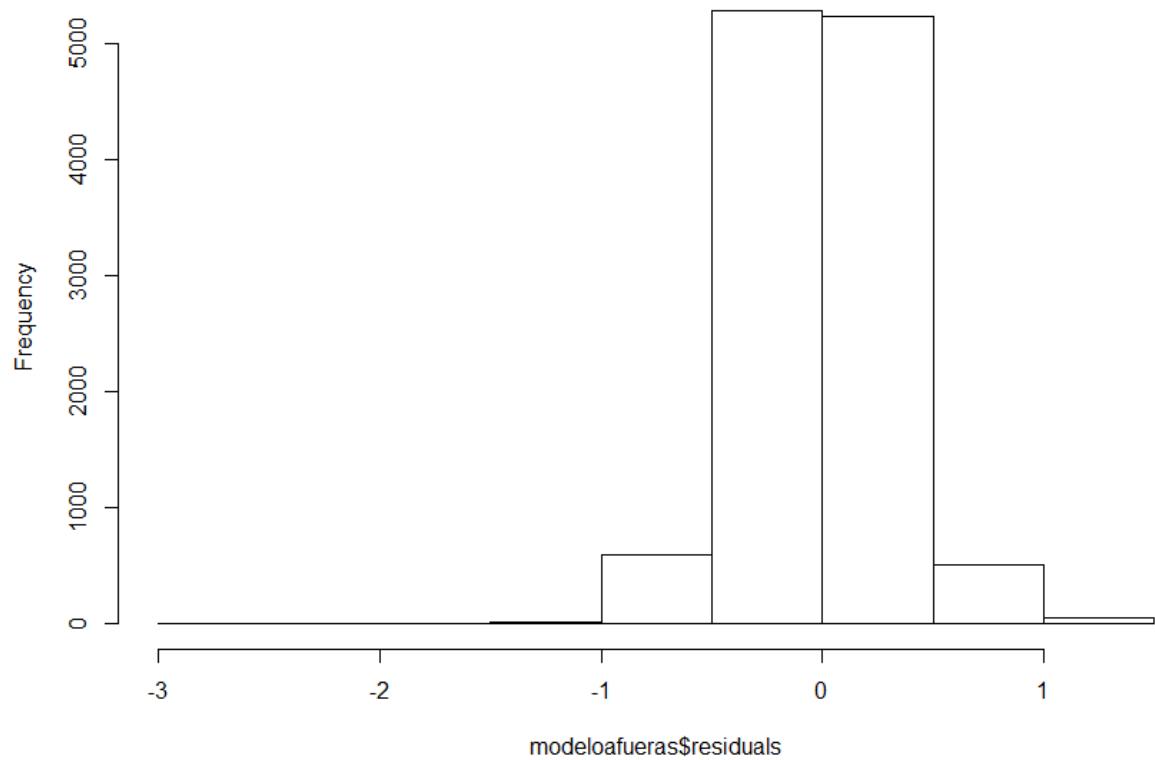
Intervalo de confianza:

	2.5 %	97.5 %
(Intercept)	1.203333e+01	1.206333e+01
sqft_living	4.012182e-04	4.143886e-04

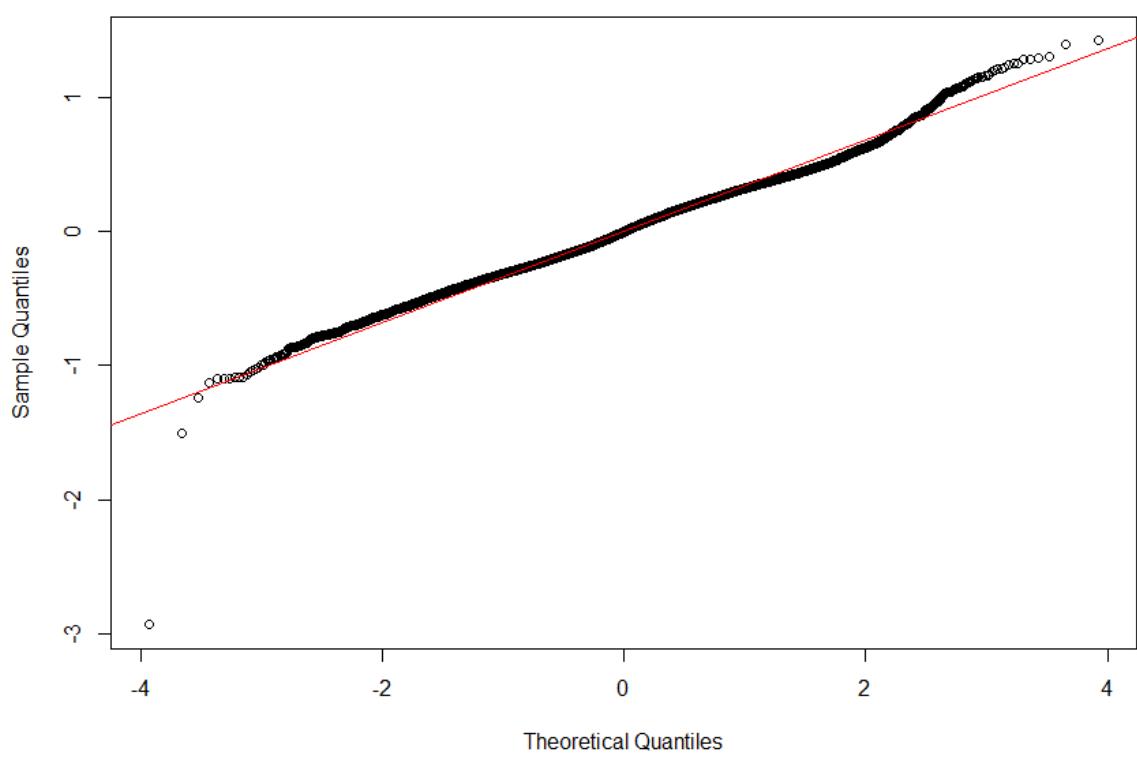
El modelo tiene error residual estándar de 0.3222.

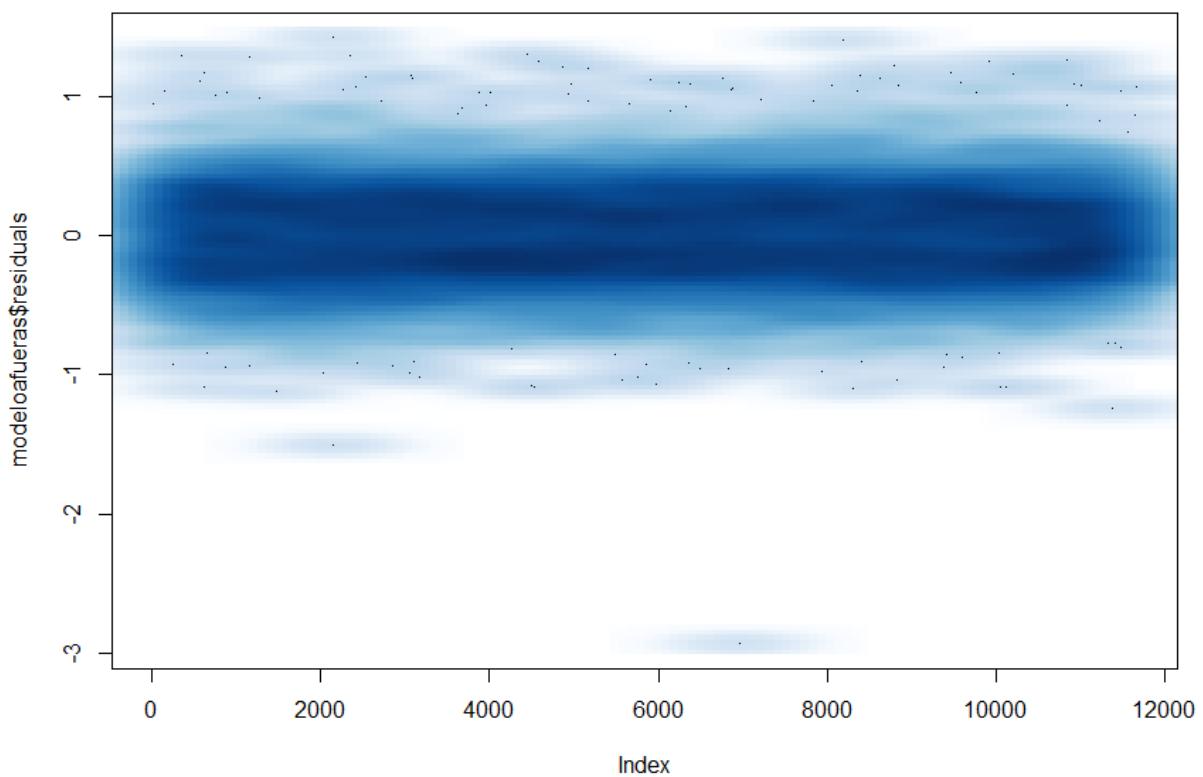
Los residuos tienen media cercana a cero y distribución parecida a normal:

Histogram of modeloafuera\$residuals



Normal Q-Q Plot





Voy a comprobar con modelo robusto si el modelo saldría mejor con un tratamiento de outliers.

```
Call: rlm(formula = log(price) ~ sqft_living, data = CASAS_AFUERAS)
Residuals:
    Min      1Q  Median      3Q     Max 
-2.972375 -0.228365 -0.003026  0.230003  1.416930 

Coefficients:
            value   Std. Error t value
(Intercept) 12.0393    0.0076 1582.1031
sqft_living  0.0004    0.0000 123.1880

Residual standard error: 0.3397 on 11680 degrees of freedom
```

Error en el modelo robusto aumenta ligeramente.

El intervalo de confianza es igual de estrecho como antes

	2.5 %	97.5 %
(Intercept)	1.202441e+01	1.205424e+01
sqft_living	4.050277e-04	4.181244e-04

Comparando la bondad de ambos modelos obtengo:

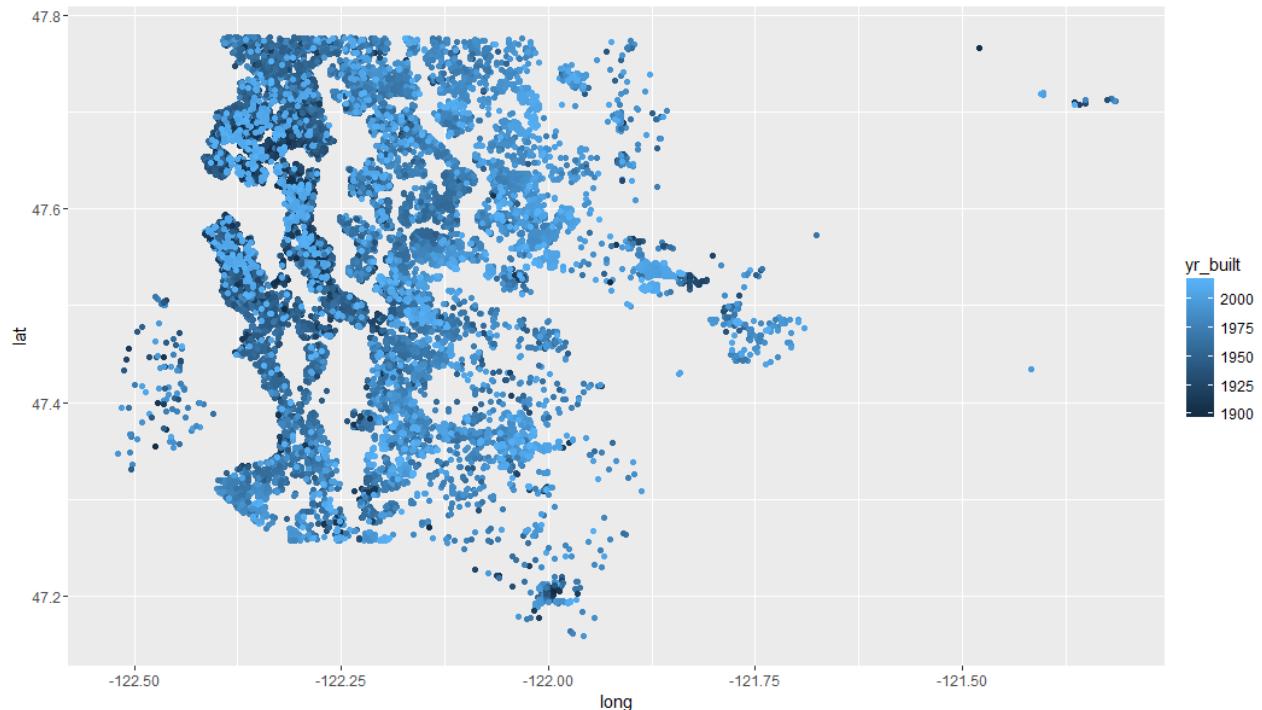
	modeloafueras	modeloafuerasrobusto
AIC	6691.435	6692.83
BIC	6713.532	6714.927

Concluyo que el modelo no mejora mediante tratamiento de outliers con una regresión robusta.

Análisis de las posibilidades de mejora de los modelos básicos mediante introducción de nuevas variables.

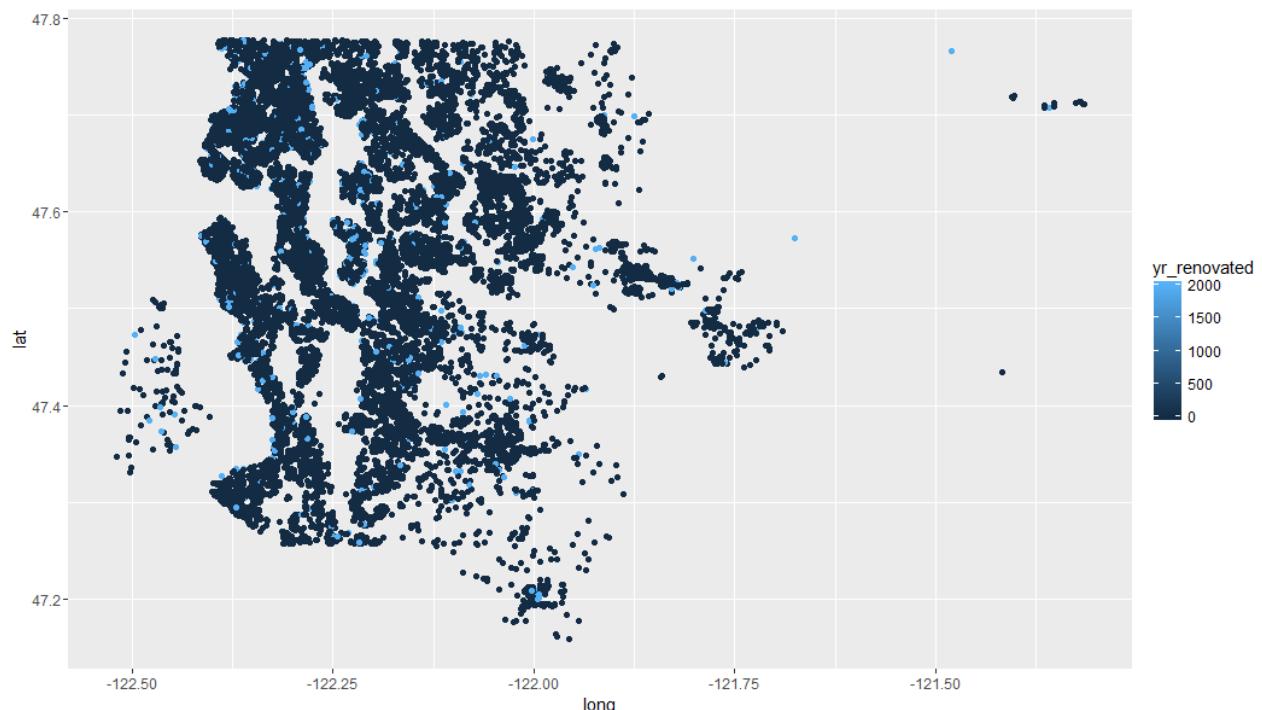
Tenemos disponibles distintas variables adicionales que podrían ayudar a mejorar el modelo básico. Voy a analizarlas:

La edad de las casas:



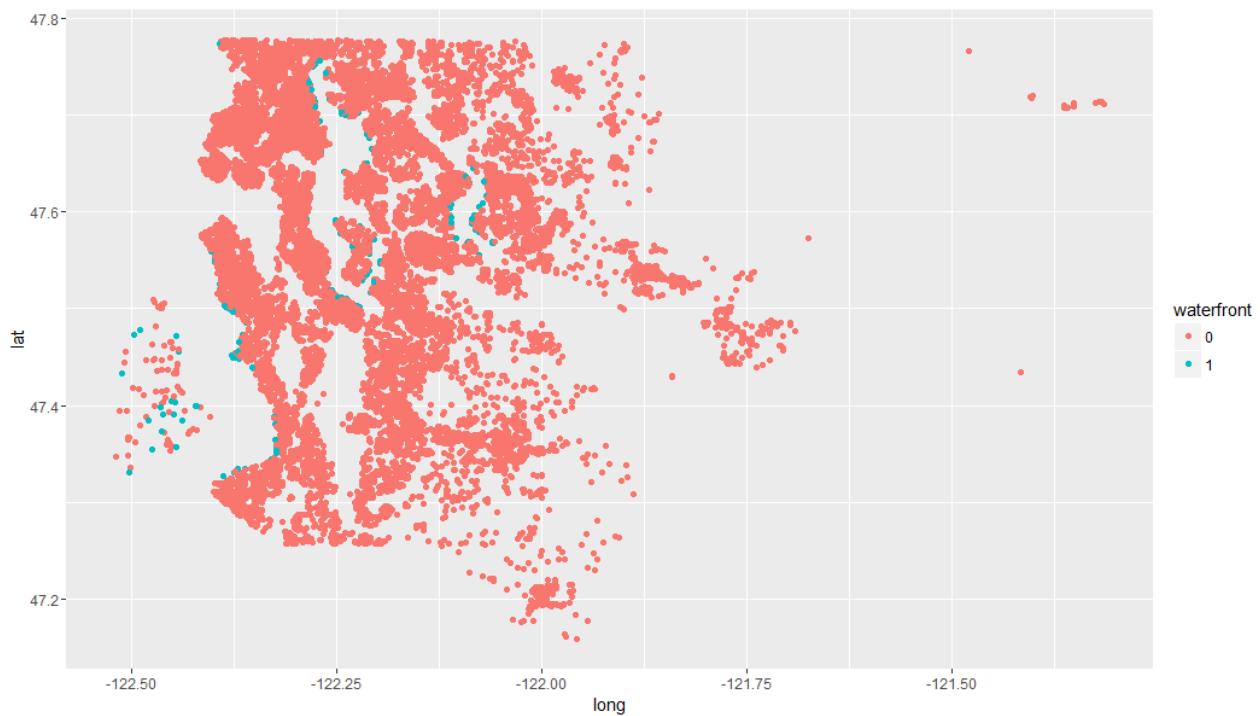
Las casas muy nuevas se encuentran en afueras, aunque también hay gran cantidad en el centro. En la cercanía a Capitol Hill se encuentran casas muy antiguas del principio del siglo XX.

Año de renovación:



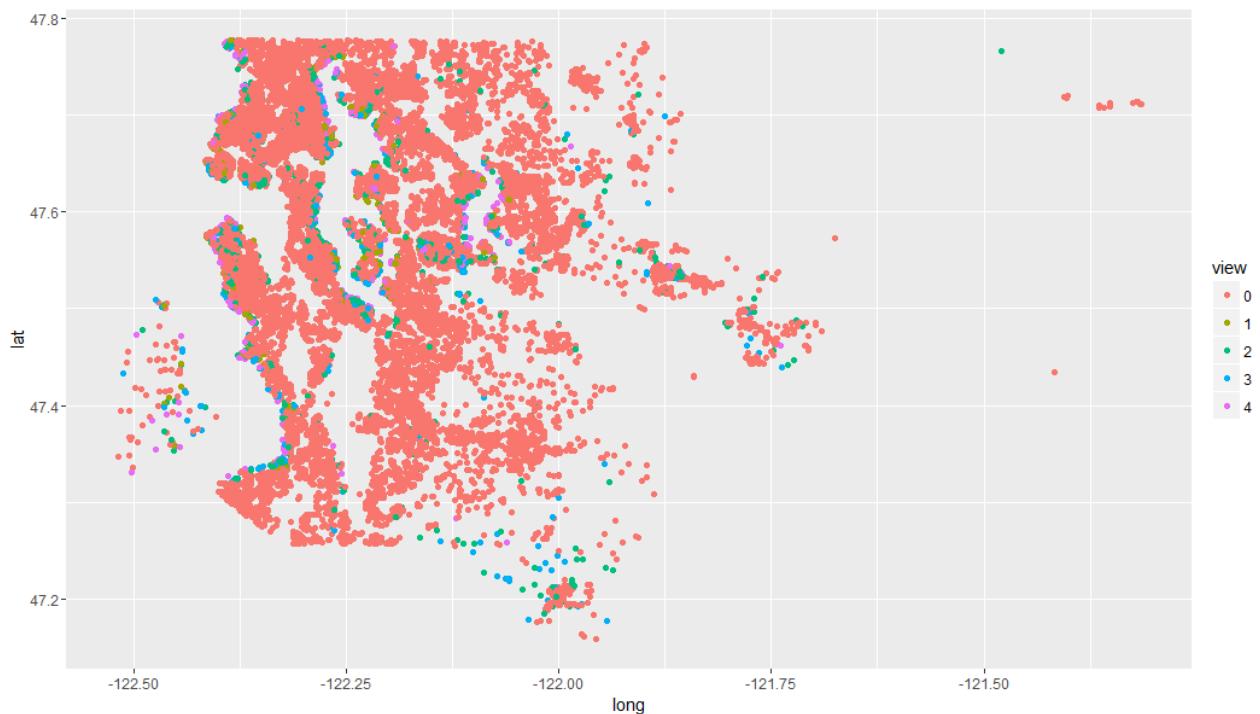
Bastantes casas en el centro y cerca del agua están renovadas. La mayoría de las casas de edad intermedia, construidas en los años 50-70 no parecen renovadas.

Vistas:



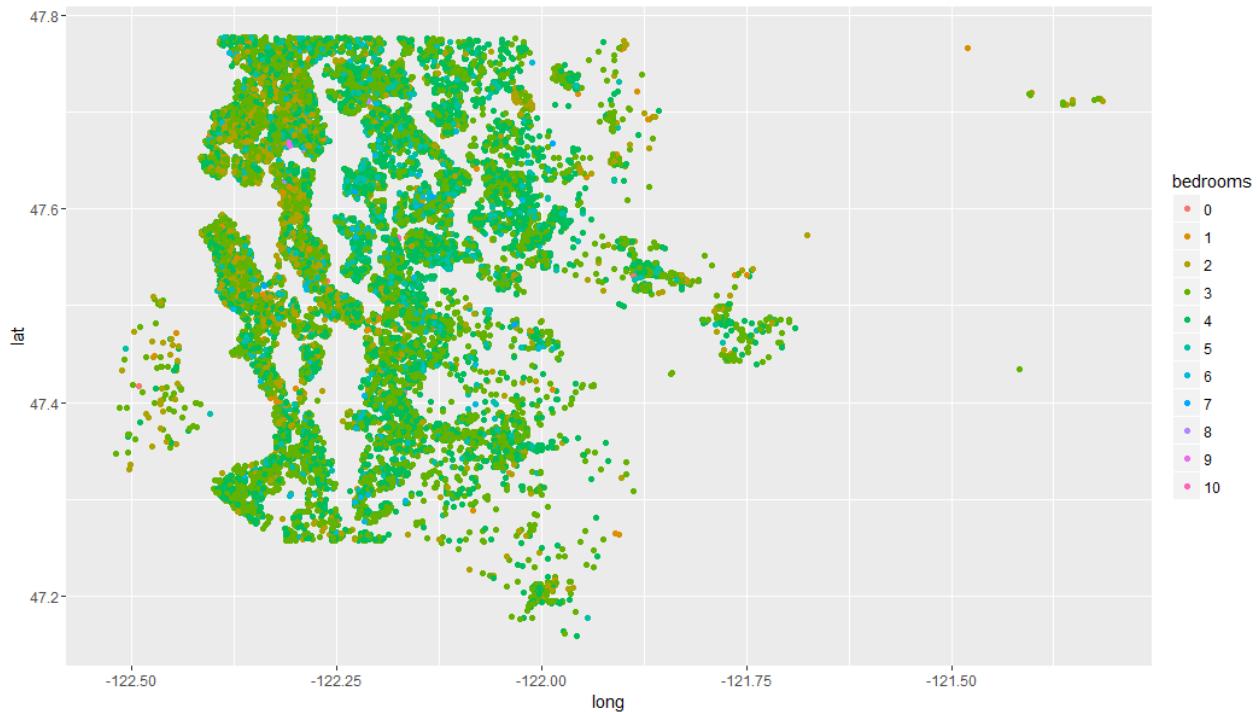
Cerca de la orilla hay casas en primera línea del mar con vistas en algunas zonas.

Orientación de la vivienda:



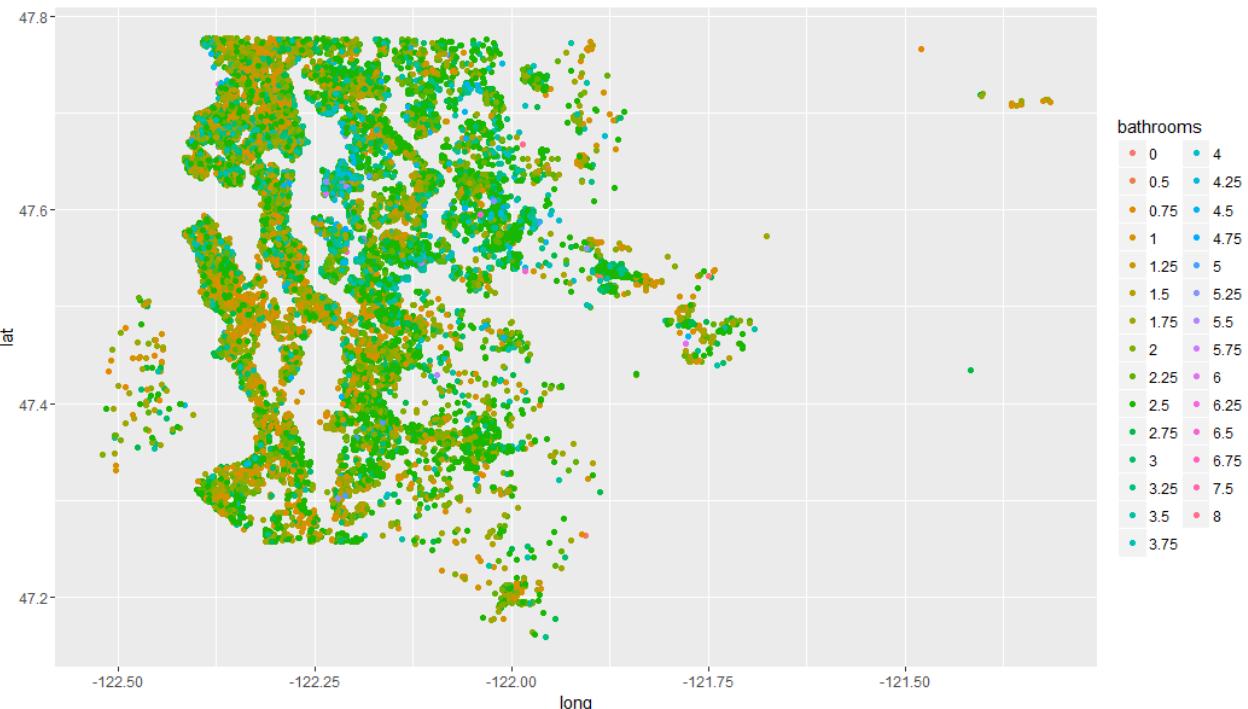
Cerca de la orilla las casas tienen también orientación mejor.

Número de habitaciones:



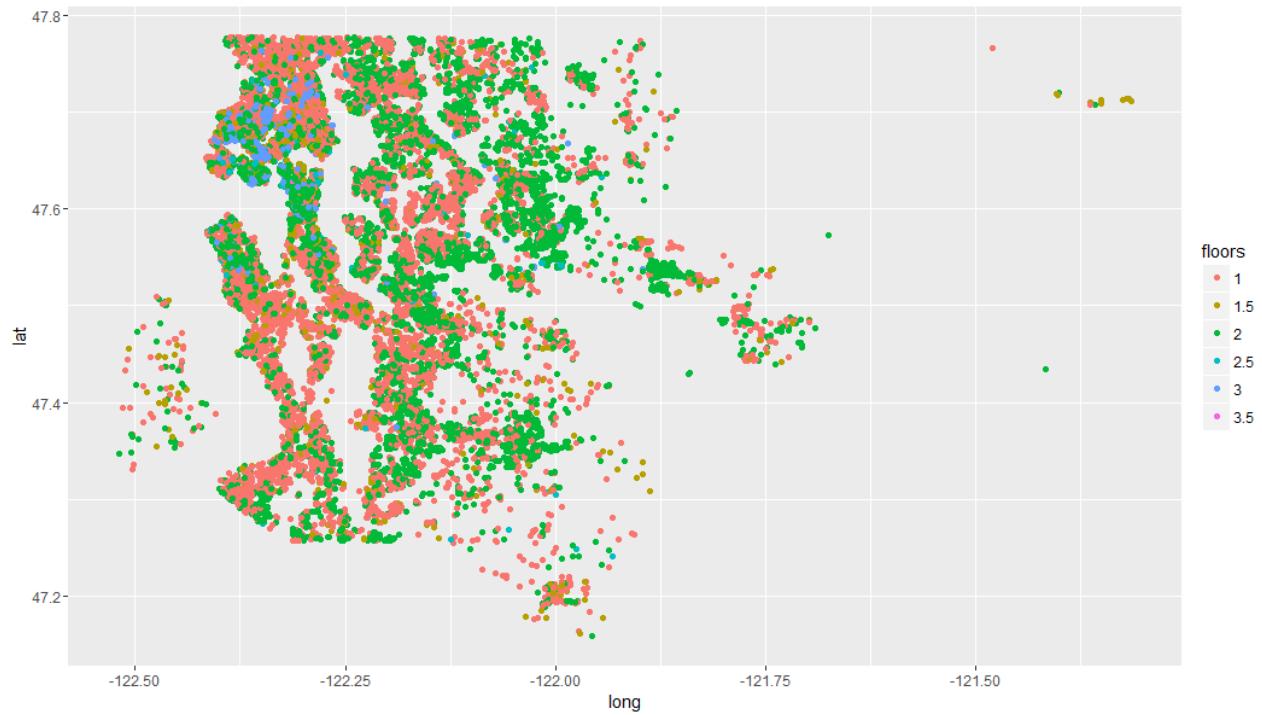
Las casas en el este parecen tener en media más habitaciones, lo cual coincide con el análisis del tamaño de las viviendas. Mayor número de 5-7 habitaciones observamos en la zona cara. También hay alguna casa de 10 habitaciones en el centro con vistas a Capitol Hill.

Número de baños:



Aunque no tengo claro que sería medio o cuarto de baño, puedo observar que las viviendas caras y grandes son donde hay más baños. Por lo contrario, las casas pequeñas tienen 1 baño o 2 como máximo por lo general.

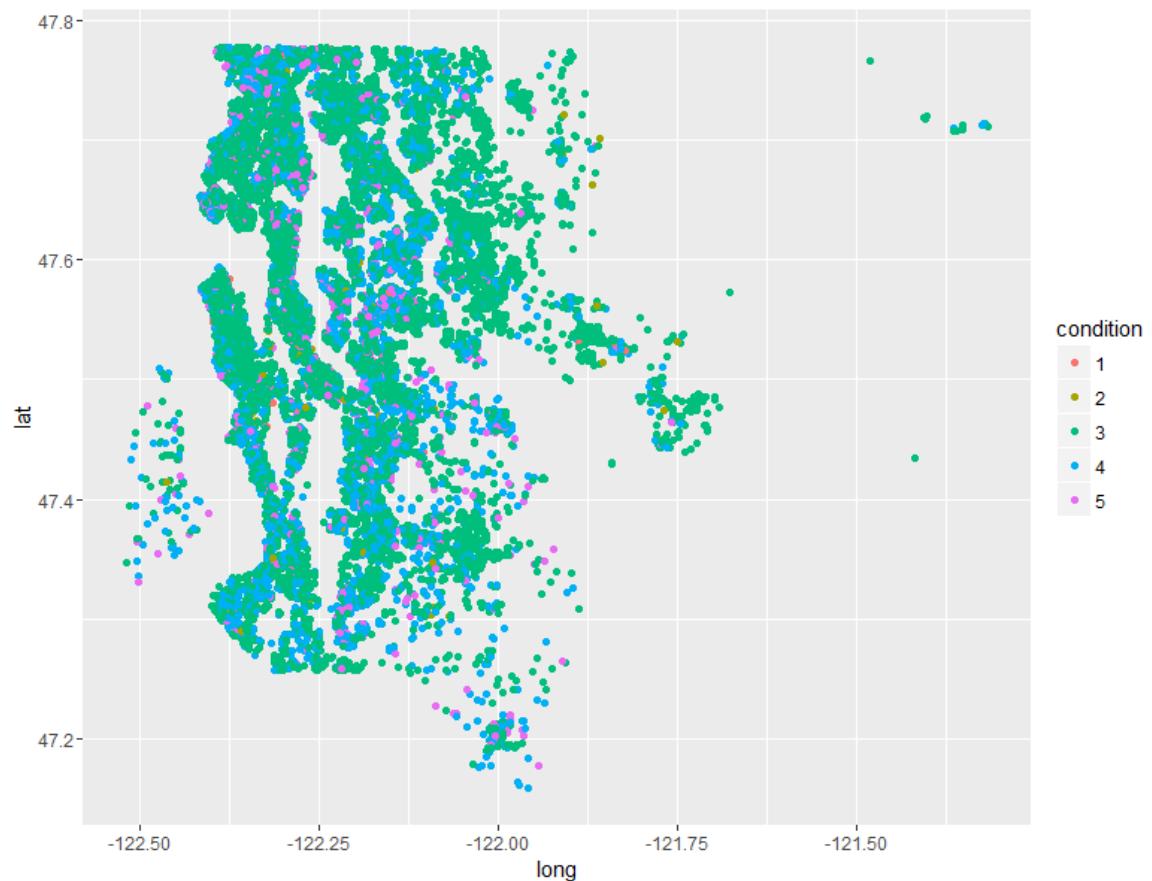
Número de plantas:



La mayoría de las viviendas tiene 1 o 2 plantas, excepto las viviendas en la zona norte de 3 plantas. Estas viviendas parecen ser las viviendas nuevas que se aprecian en el gráfico anterior.

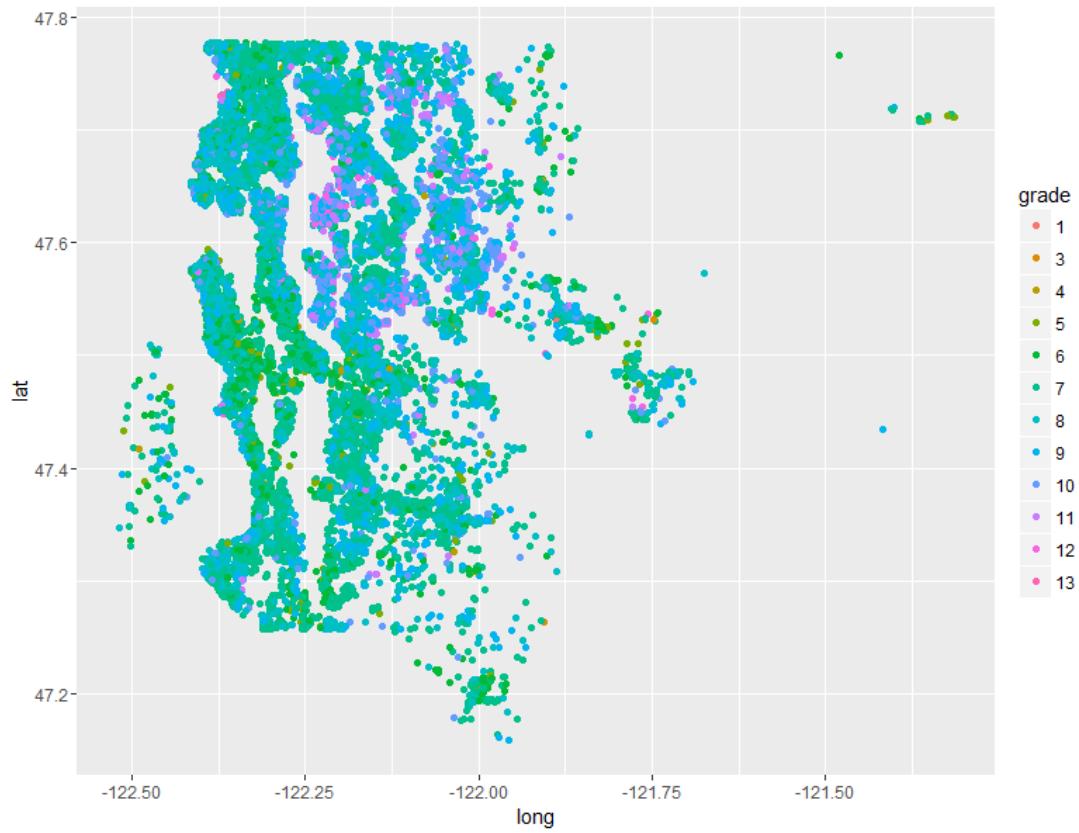
Las viviendas de 2 plantas en otras zonas también parecen ser más nuevas que las de 1 planta en la zona oeste.

Estado de la vivienda:



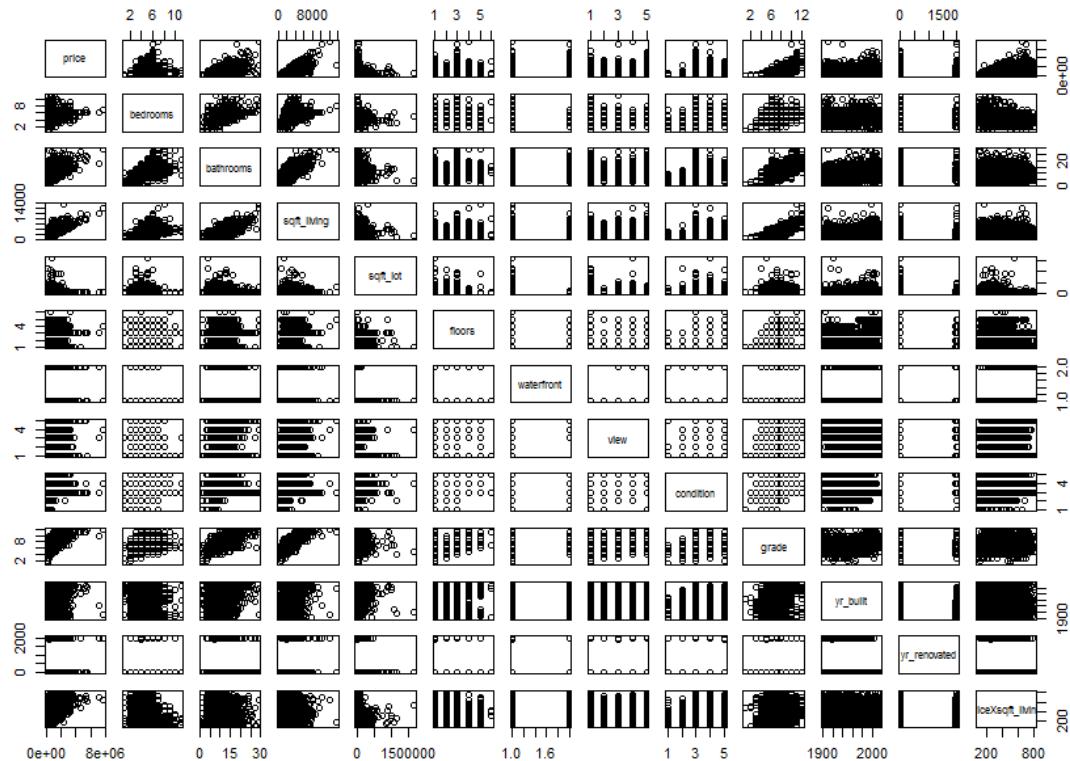
La mayoría de las viviendas está en medio o buen estado. Las casas peor mantenidas prevalecen en afuera.

Diseño:



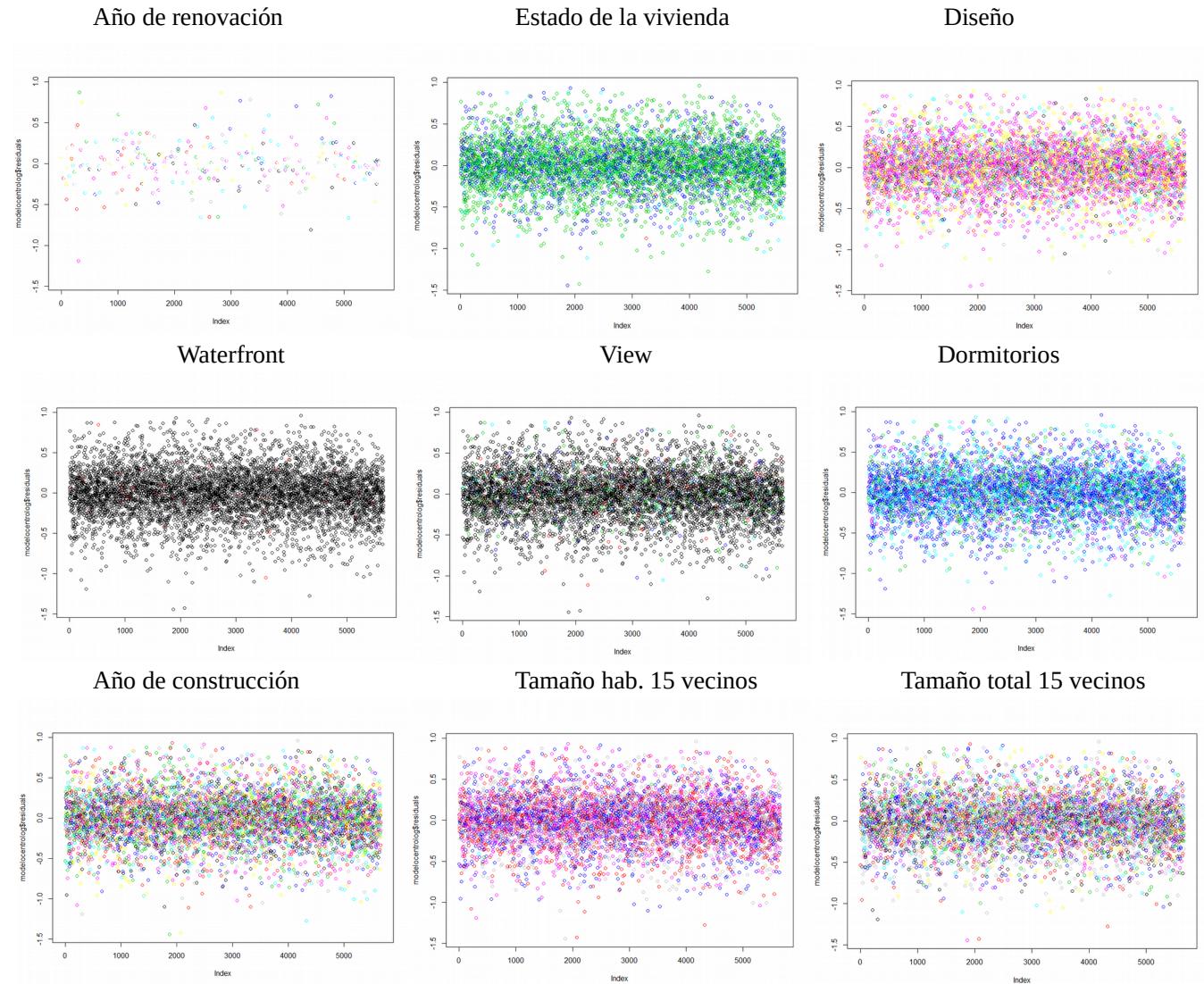
Las casas mejor diseñadas se encuentran definitivamente en la parte noreste, incluyendo los barrios caros y las zonas de viviendas nuevas. Las casas más sencillas están en la parte oeste de la ciudad.

Veamos la relación entre las variables de forma general:



Parece que cuanto más grande la vivienda (sqft_living) mas cara. La superficie parece tener relación también con el numero de dormitorios y de los baños. Cuanto mejor el diseño, mas baños y dormitorios por lo general. El precio por metro cuadrado es mayor cuando el diseño y estado de la casa mejora.

Voy a revisar si los residuos del modelo básico para las casas en el centro tienen algún patrón relacionado con estas variables.



No se ven patrones muy definidos aunque puede haber pequeñas asimetrías en el reparto de los residuos y la introducción de más variables puede ayudar a reducir el sesgo del modelo.

Veamos si con variables nuevas el modelo mejoraría. Construyo siguientes modelos añadiendo variables de firma incremental:

	AIC	BIC
Modelo básico	2181.165	2201.111
Añadiendo variable grade	1365.578	1445.361
Añadiendo variable condition	1072.312	1178.69
Añadiendo variable water front	986.9804	1100.006
Añadiendo variable view	841.8923	981.5124
Añadiendo var. año de constr.	303.2712	449.5398
Añadiendo var. sqft_living15	9.768881	162.686

Me quedo con el modelo con AIC y BIC más bajo es el modelo que incluye las variables: sqft_living, grade, condition, waterfront, view, yr_built, sqft_living15.

El resultado parece bastante lógico: se paga más por cada unidad adicional de superficie en función del estado y diseño de la vivienda, dependiendo si tiene vistas y orientación mejor y también, dado que estamos hablando de la zona céntrica y cara que tiene edificios muy antiguos, en función de la edad de edificio y el vecindario que la rodea.

El mejor modelo para las casas en el centro que he conseguido es:

```
lm(formula = log(price) ~ sqft_living + grade + condition + waterfront +
view + yr_built + sqft_living15, data = CASAS_CENTRO)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.33200 -0.14814  0.00627  0.15238  1.00673 

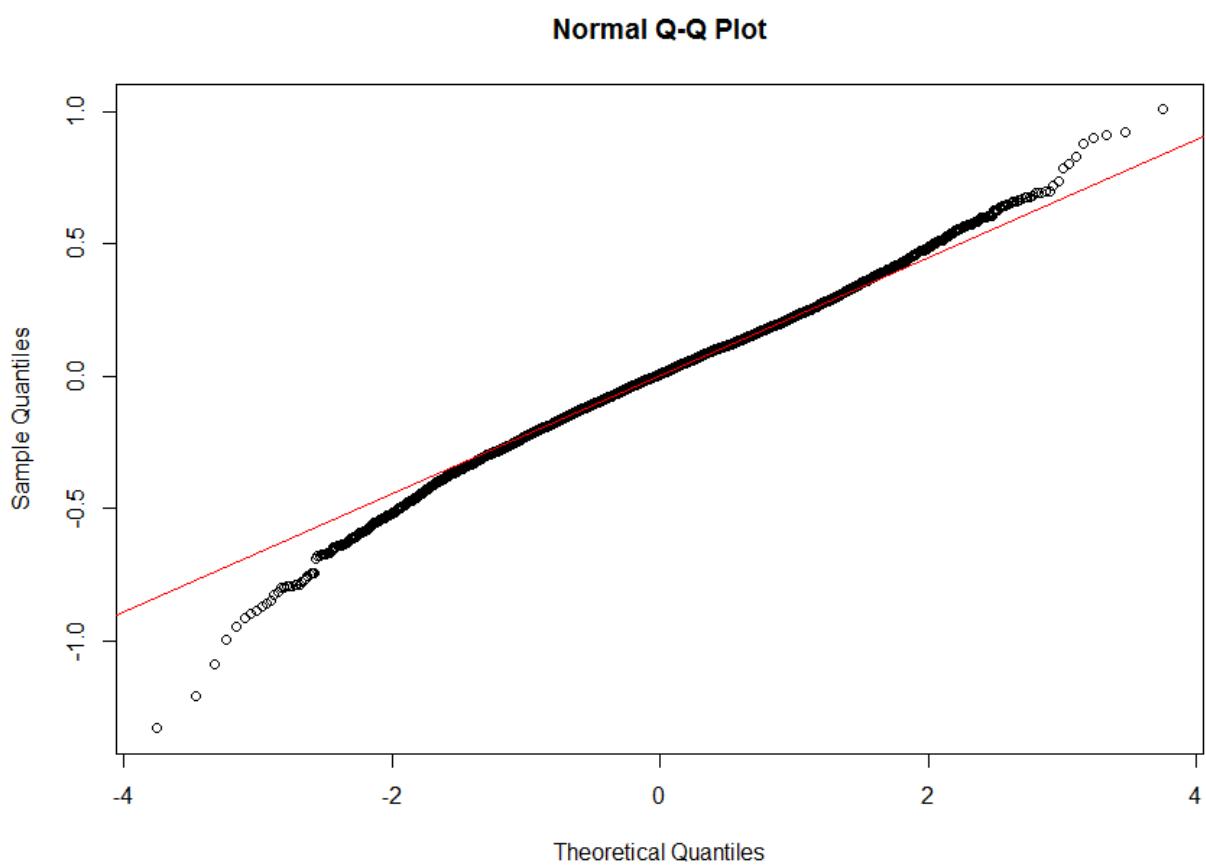
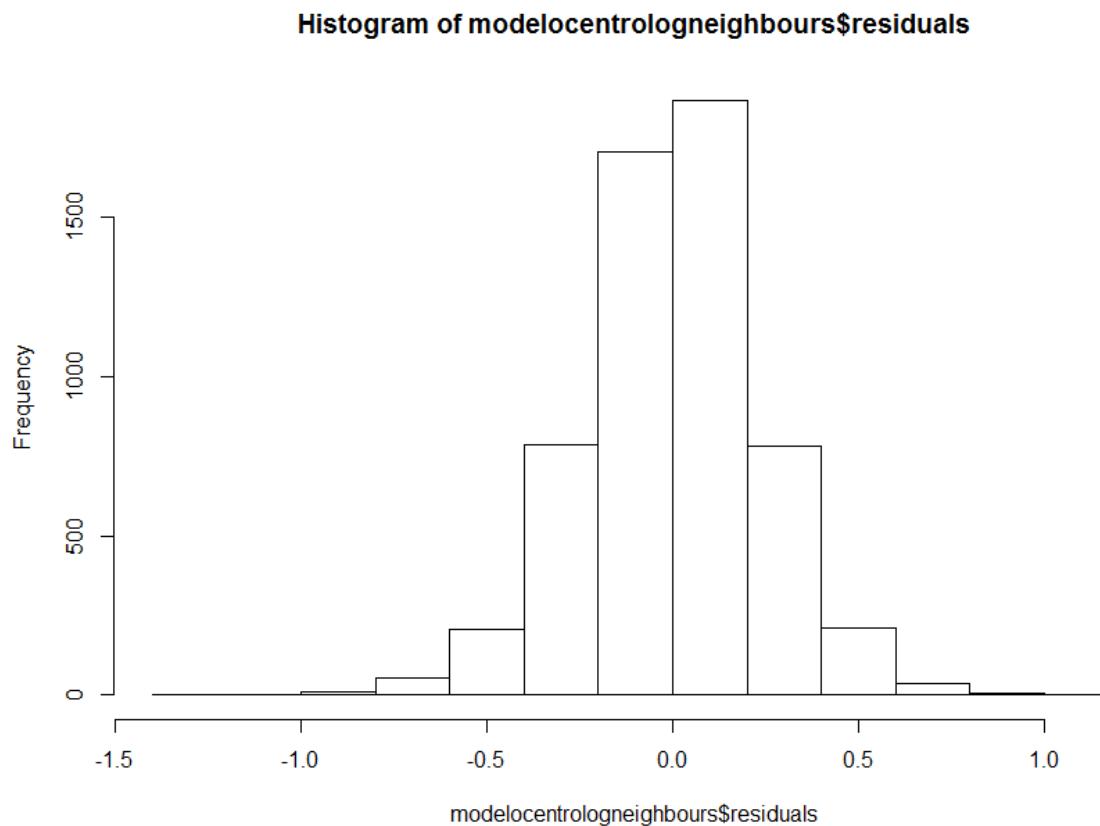
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.695e+01  2.757e-01 61.495 < 2e-16 ***
sqft_living 1.956e-04  5.692e-06 34.370 < 2e-16 ***
grade5      3.220e-01  1.470e-01  2.191  0.028521 *  
grade6      3.638e-01  1.403e-01  2.594  0.009513 ** 
grade7      5.761e-01  1.399e-01  4.117  3.90e-05 *** 
grade8      7.334e-01  1.401e-01  5.235  1.71e-07 *** 
grade9      9.411e-01  1.405e-01  6.697  2.34e-11 *** 
grade10     1.074e+00  1.414e-01  7.600  3.43e-14 *** 
grade11     1.123e+00  1.430e-01  7.851  4.89e-15 *** 
grade12     1.246e+00  1.484e-01  8.398 < 2e-16 *** 
grade13     1.101e+00  1.707e-01  6.447  1.23e-10 *** 
condition2  8.361e-02  7.948e-02  1.052  0.292873  
condition3  2.432e-01  7.122e-02  3.415  0.000643 *** 
condition4  2.774e-01  7.125e-02  3.893  0.000100 *** 
condition5  3.140e-01  7.166e-02  4.382  1.20e-05 *** 
waterfront1 2.762e-01  5.244e-02  5.266  1.45e-07 *** 
view1       7.377e-02  2.167e-02  3.404  0.000669 *** 
view2       4.137e-02  1.387e-02  2.981  0.002882 ** 
view3       6.906e-02  2.124e-02  3.251  0.001155 ** 
view4       1.925e-01  2.909e-02  6.619  3.96e-11 *** 
yr_built    -2.665e-03  1.161e-04 -22.952 < 2e-16 *** 
sqft_living15 1.269e-04  7.303e-06 17.382 < 2e-16 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2417 on 5680 degrees of freedom
```

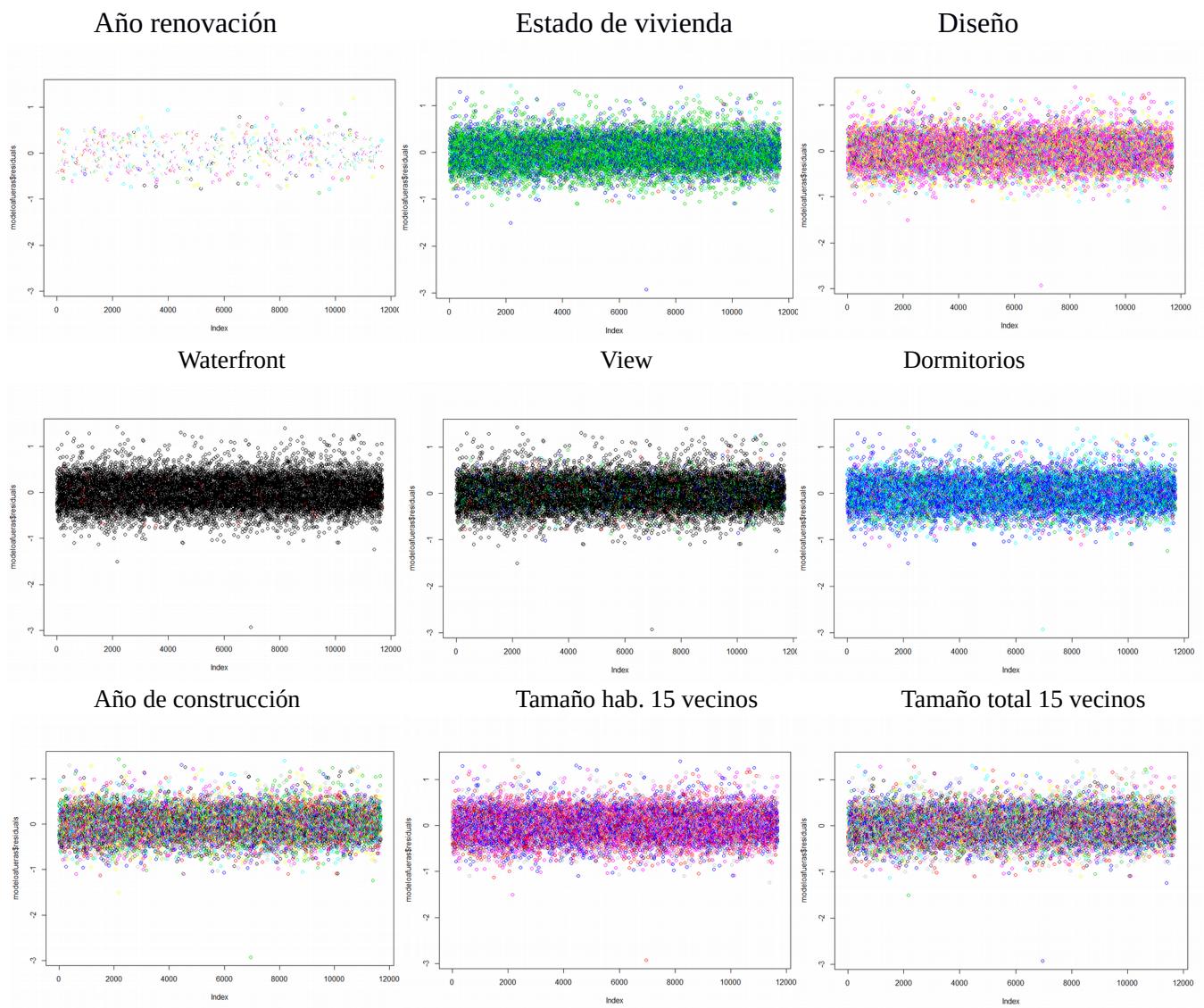
Con el siguiente intervalo de confianza:

	2.5 %	97.5 %
(Intercept)	16.4124539244	17.4933343544
sqft_living	0.0001844641	0.0002067795
grade5	0.0338455599	0.6102337232
grade6	0.0888631426	0.6387887735
grade7	0.3017652366	0.8504760830
grade8	0.4587676497	1.0080442668
grade9	0.6656232047	1.2166284838
grade10	0.7972589519	1.3514841629
grade11	0.8423398500	1.4029797960
grade12	0.9553637004	1.5371965406
grade13	0.7660467502	1.4354376230
condition2	-0.0722034453	0.2394198254
condition3	0.1035835269	0.3828214714
condition4	0.1376887838	0.4170541500
condition5	0.1735159923	0.4544788298
waterfront1	0.1733531980	0.3789726036
view1	0.0312813888	0.1162582106
view2	0.0141662530	0.0685666002
view3	0.0274197494	0.1106985585
view4	0.1354910671	0.2495292921
yr_built	-0.0028928034	-0.0024375277
sqft_living15	0.0001126247	0.0001412588

Al añadir mas variables ha bajado el sesgo del modelo pero ha aumentado la dispersión.



Revisando los residuos del modelo básico de las casas en afueras:



No destacan patrones muy visibles. Aún así probaré mejorar el modelo con algunas variables adicionales y comprobando los resultados con AIC y BIC.

	AIC	BIC
Modelo básico	6691.435	6713.532
Añadiendo variable grade	4728.72	4831.841
Añadiendo variable condition	4564.302	4696.887
Añadiendo variable water front	3912.455	4052.405
Añadiendo variable view	3567.193	3736.606
Añadiendo variable yr_build	3500.594	3677.373
Añadiendo var. sqft_living15	3116.203	3300.348

El modelo con AIC y BIC más bajo incluye adicionalmente a la superficie las variables grade, condition, water front, view, yr_build y sqft_living15.

El mejor modelo para las casas de las afueras que he conseguido es:

```
lm(formula = log(price) ~ sqft_living + grade + condition + waterfront +
view + yr_built + sqft_living15, data = CASAS_AFUERAS)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.30240 -0.19647  0.00895  0.19840  1.26175 

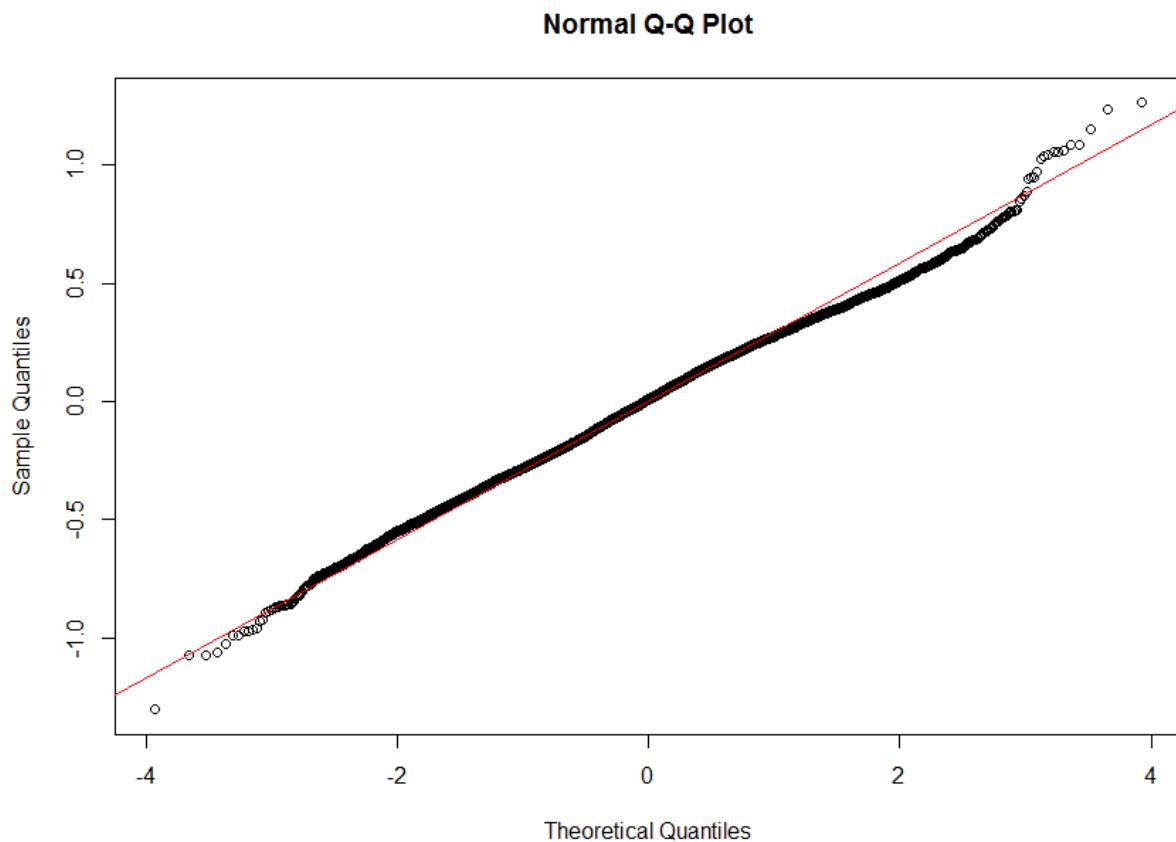
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.464e+01  3.970e-01 36.883 < 2e-16 ***
sqft_living 1.733e-04  5.281e-06 32.821 < 2e-16 ***
grade3      4.186e-03  3.299e-01  0.013 0.989879  
grade4      -9.394e-02  2.981e-01 -0.315 0.752633  
grade5      3.879e-02  2.888e-01  0.134 0.893159  
grade6      1.714e-01  2.887e-01  0.594 0.552721  
grade7      3.554e-01  2.888e-01  1.231 0.218372  
grade8      5.386e-01  2.888e-01  1.865 0.062260 .  
grade9      6.851e-01  2.890e-01  2.370 0.017789 *  
grade10     8.117e-01  2.893e-01  2.806 0.005029 ** 
grade11     8.821e-01  2.900e-01  3.041 0.002360 ** 
grade12     9.266e-01  2.926e-01  3.166 0.001547 ** 
grade13     1.188e+00  3.158e-01  3.762 0.000170 *** 
condition2  5.141e-02  8.865e-02  0.580 0.562006  
condition3  2.172e-01  8.409e-02  2.583 0.009797 ** 
condition4  2.227e-01  8.410e-02  2.648 0.008117 ** 
condition5  3.101e-01  8.456e-02  3.668 0.000246 *** 
waterfront1 4.834e-01  3.601e-02 13.423 < 2e-16 ***
view1       1.920e-01  2.457e-02  7.814 6.02e-15 *** 
view2       1.055e-01  1.348e-02  7.825 5.54e-15 *** 
view3       1.525e-01  1.802e-02  8.464 < 2e-16 *** 
view4       2.699e-01  2.895e-02  9.324 < 2e-16 *** 
yr_builtin  -1.550e-03  1.459e-04 -10.624 < 2e-16 *** 
sqft_living15 1.312e-04  6.628e-06 19.800 < 2e-16 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2762 on 11658 degrees of freedom
```

Con el siguiente intervalo de confianza:

	2.5 %	97.5 %
(Intercept)	13.8644725009	15.4208591015
sqft_living	0.0001629813	0.0001836853
grade3	-0.6425401547	0.6509111722
grade4	-0.6781940499	0.4903093378
grade5	-0.5273143004	0.6048936775
grade6	-0.3945411572	0.7373846690
grade7	-0.2105731534	0.9214640907
grade8	-0.0275998845	1.1047757292
grade9	0.1185431032	1.2516288575
grade10	0.2446080727	1.3787181997
grade11	0.3135838437	1.4505774201
grade12	0.3529985006	1.5001942400
grade13	0.5689625383	1.8070244083
condition2	-0.1223648907	0.2251798827
condition3	0.0524025494	0.3820597821
condition4	0.0578114035	0.3874972703
condition5	0.1443791991	0.4758818881
waterfront1	0.4127877733	0.5539627490
view1	0.1438250043	0.2401475609
view2	0.0790639646	0.1319184373
view3	0.1172169597	0.1878704047
view4	0.2131836635	0.3266753942
yr_builtin	-0.0018355517	-0.0012637241
sqft_living15	0.0001182366	0.0001442191

Al añadir mas variables ha bajado el sesgo del modelo pero ha aumentado la dispersión.



Uno solo modelo con variable dummy:

Puedo juntar los dos modelos en uno solo, usando una variable dummy:

```
lm(formula = log(price) ~ sqft_living + grade + condition + waterfront +
  view + yr_builtin + sqft_living15 + centro, data = CASAS_TODAS)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.41488	-0.17958	0.00751	0.18192	1.24514

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.597e+01	3.221e-01	49.594	< 2e-16 ***
sqft_living	1.853e-04	3.922e-06	47.247	< 2e-16 ***
grade3	-3.026e-02	3.126e-01	-0.097	0.922886
grade4	-1.637e-01	2.797e-01	-0.585	0.558289
grade5	3.203e-03	2.723e-01	0.012	0.990616
grade6	1.272e-01	2.721e-01	0.467	0.640316
grade7	3.247e-01	2.721e-01	1.193	0.232780
grade8	4.982e-01	2.722e-01	1.830	0.067229
grade9	6.653e-01	2.723e-01	2.443	0.014568 *
grade10	7.888e-01	2.725e-01	2.895	0.0003800 **
grade11	8.501e-01	2.730e-01	3.114	0.001846 **
grade12	9.143e-01	2.744e-01	3.332	0.000865 ***
grade13	9.772e-01	2.839e-01	3.442	0.0000578 ***
condition2	7.414e-02	6.041e-02	1.227	0.219688
condition3	2.405e-01	5.615e-02	4.283	1.86e-05 ***
condition4	2.536e-01	5.615e-02	4.517	6.32e-06 ***
condition5	3.212e-01	5.647e-02	5.688	1.30e-08 ***
waterfront1	4.519e-01	2.878e-02	15.698	< 2e-16 ***
view1	1.317e-01	1.673e-02	7.873	3.65e-15 ***
view2	7.586e-02	9.865e-03	7.690	1.55e-14 ***
view3	1.183e-01	1.390e-02	8.509	< 2e-16 ***
view4	2.356e-01	2.094e-02	11.248	< 2e-16 ***
yr_builtin	-2.230e-03	9.231e-05	-24.159	< 2e-16 ***
sqft_living15	1.318e-04	4.888e-06	26.959	< 2e-16 ***
centro1	3.914e-01	4.731e-03	82.730	< 2e-16 ***

Obtengo el siguiente intervalo de confianza bastante amplio:

```
confint(modelotodasneihgbours, level=0.95)
2.5 % 97.5 %
(Intercept) 15.3425625290 16.6052211388
sqft_living 0.0001776162 0.0001929913
grade3 -0.6429593181 0.5824426237
grade4 -0.7119215282 0.3844757786
grade5 -0.5305711599 0.5369766563
grade6 -0.4062252940 0.6605267968
grade7 -0.2086911439 0.8581832388
grade8 -0.0353439928 1.0317511969
grade9 0.1315556886 1.1991369223
grade10 0.2546765604 1.3229693315
grade11 0.3150963593 1.3851420333
grade12 0.3763855098 1.4522282943
grade13 0.4207609162 1.5335917752
condition2 -0.0442607470 0.1925501584
condition3 0.1304001558 0.3505092741
condition4 0.1435747547 0.3637114873
condition5 0.2105151447 0.4318810182
waterfront1 0.3954311186 0.5082732352
view1 0.0989281288 0.1645116521
view2 0.0565254334 0.0951995744
view3 0.0910440981 0.1455463085
view4 0.1945038457 0.2765964586
yr_built -0.0024109570 -0.0020491016
sqft_living15 0.0001222082 0.0001413720
centro 0.3821327353 0.4006797109
```

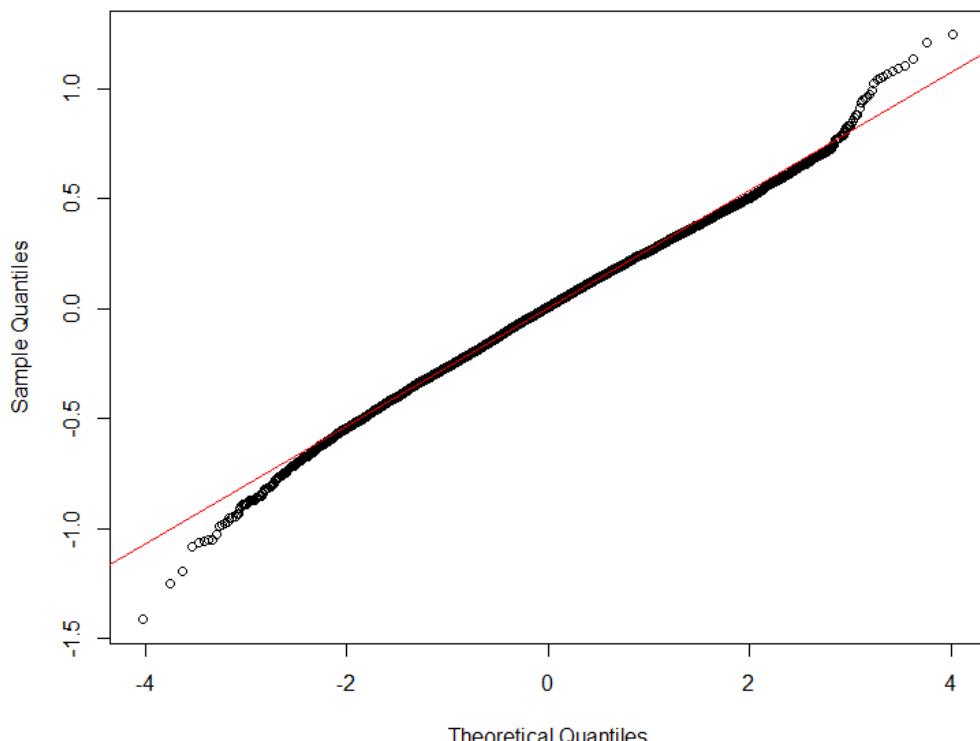
Puedo decir que con nivel de confianza de 95% una subida de superficie en 1 pie cuadrado de una vivienda de mil pies, construida en el año 2000 al lado de otras viviendas de 1000 pies, con grade 13, condition 5, con vista al mar, view 4 y en el centro, está asociada con un aumento de precio entre 44.16886 y 1477.939 dólares.

$$\exp(15.3425625290 + 1001 * 0.0001776162 + 0.4207609162 + 0.2105151447 + 0.3954311186 + 0.1945038457 + 2000 * (-0.0024109570) + 1000 * 0.0001222082 + 0.3821327353) - \exp(15.3425625290 + 1000 * 0.0001776162 + 0.4207609162 + 0.2105151447 + 0.3954311186 + 0.1945038457 + 2000 * (-0.0024109570) + 1000 * 0.0001222082 + 0.3821327353) = 44.16886$$

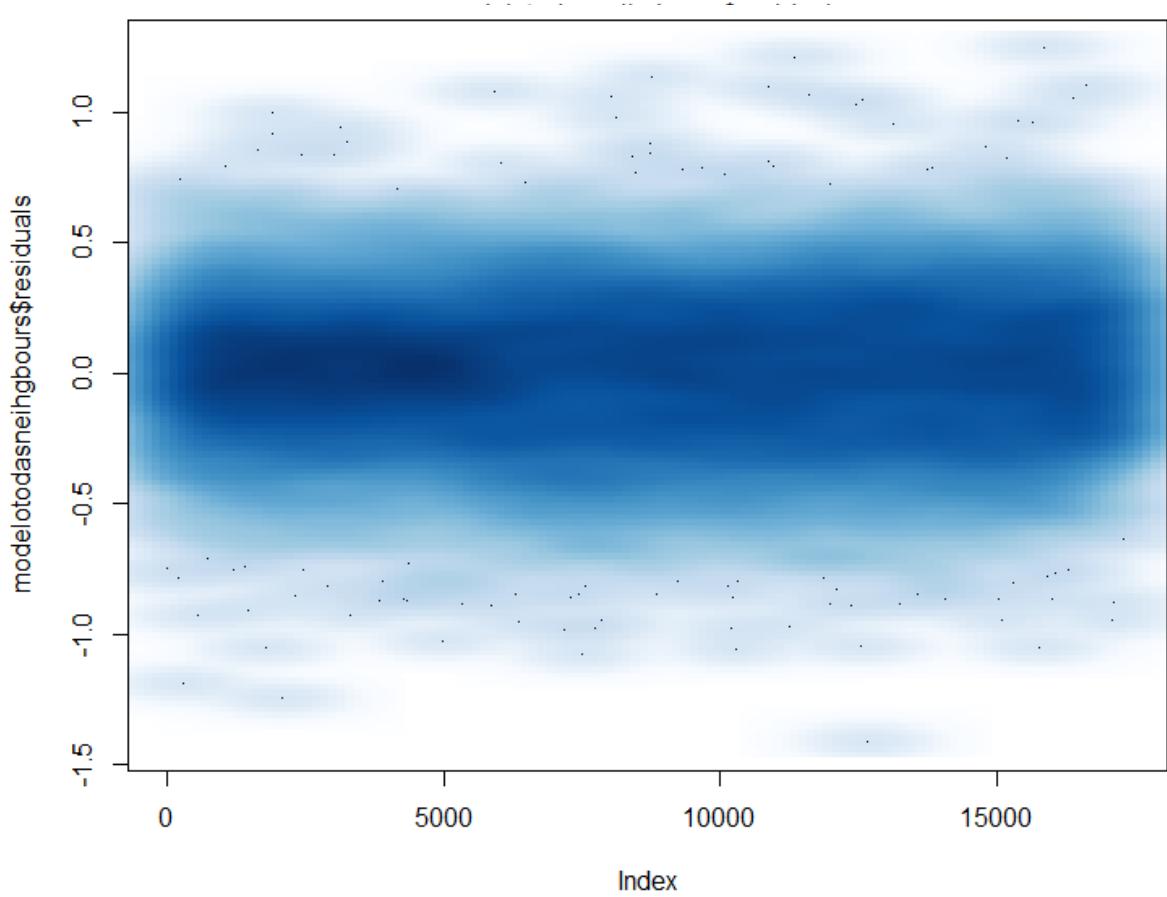
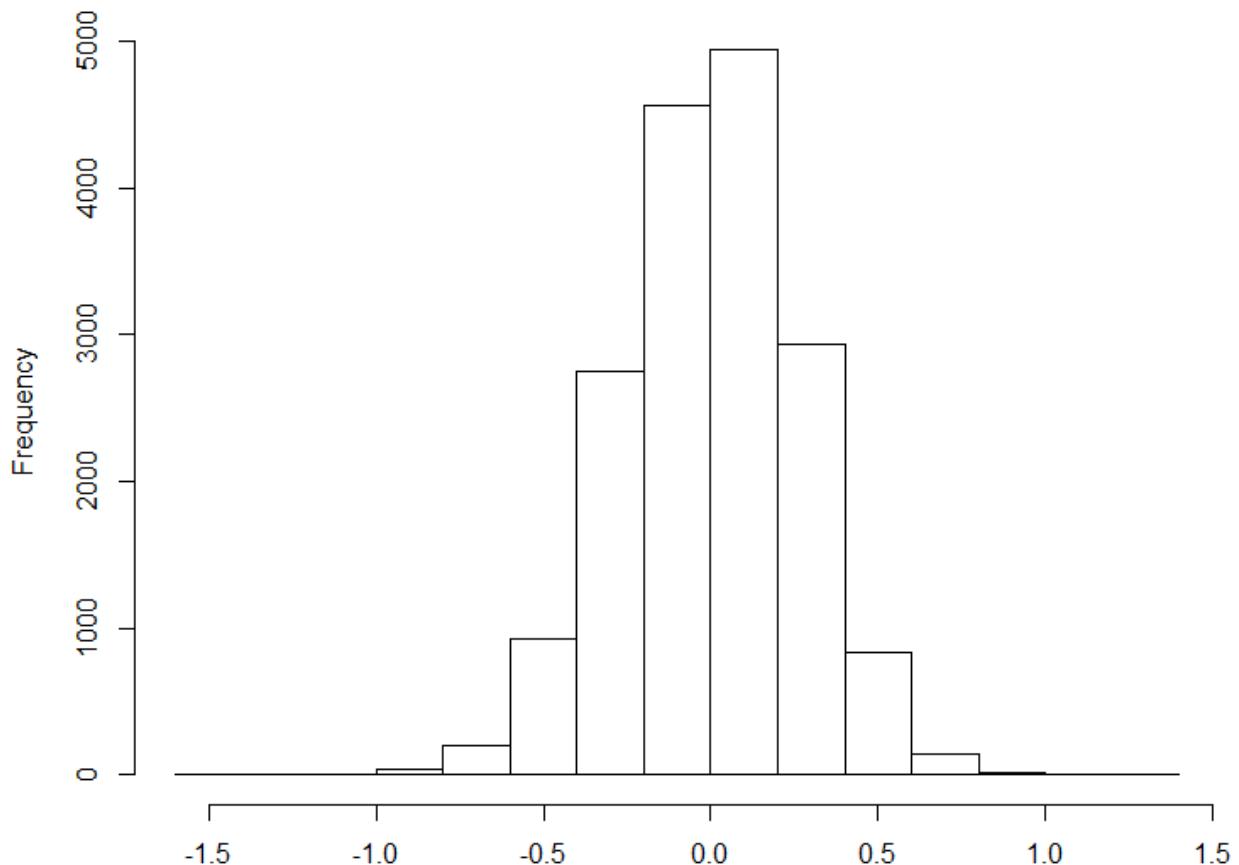
$$\exp(16.6052211388 + 1001 * 0.0001929913 + 1.5335917752 + 0.4318810182 + 0.5082732352 + 0.2765964586 + 2000 * (-0.0020491016) + 0.0001413720 + 0.4006797109) - \exp(16.6052211388 + 1000 * 0.0001929913 + 1.5335917752 + 0.4318810182 + 0.5082732352 + 0.2765964586 + 2000 * (-0.0020491016) + 0.0001413720 + 0.4006797109) = 1477.939$$

Los residuos tienen distribución parecida a normal, con algo de distorsión en las colas.

Normal Q-Q Plot



Histogram of modelotodasneihgbours\$residuals



Estimación el precio de venta de los inmuebles de la cartera de la empresa.

En esta parte voy a construir el modelo predictivo para poder estimar el precio de las viviendas del data set de test.

Divido la población en la parte de entrenamiento y parte de test y construyo un modelo de regresión lineal sobre la muestra de entrenamiento. Voy a hacer un modelo conjunto para las casas del centro y afuera, incluyendo la dummy “centro”.

Comprobación multicolinealidad

Voy a realizar la primera comprobación para ver si puedo incluir las variables elegidas sin provocar multicolinealidad:

Martiz de correlaciones para las variables numéricas:

	price	sqft_living	sqft_lot	sqft_above	sqft_basement	yr_built	yr_renovated	sqft_living15	sqft_lot15	
priceXsqft_living	1.00000000	0.70291635	0.08823811	0.60527752	0.33122956	0.05252215	0.123502403	0.583480821	0.080806426	
0.55452985	0.70291635	1.00000000	0.16696728	0.87631944	0.44288611	0.31595847	0.053794393	0.756274241	0.178306444	
-0.09013088	0.08823811	0.16696728	1.00000000	0.17600546	0.01869188	0.05150568	0.007099190	0.147707827	0.727774079	
-0.03124128	0.60527752	0.87631944	0.17600546	1.00000000	-0.04379916	0.42350762	0.021792976	0.732554007	0.188503973	
-0.08839251	0.33122956	0.44288611	0.01869188	-0.04379916	1.00000000	-0.13296294	0.071001674	0.205004207	0.018945951	
-0.02240592	0.05252215	0.31595847	0.05150568	0.42350762	-0.13296294	1.00000000	-0.222075780	0.327035852	0.072510500	
-0.29121236	0.12350240	0.05379439	0.00709919	0.02179298	0.07100167	-0.22207578	1.00000000	-0.003330807	0.008415829	
0.10123502	0.58348082	0.75627424	0.14770783	0.73255401	0.20500421	0.32703585	-0.003330807	1.00000000	0.184561578	
0.03924912	0.08080643	0.17830644	0.72777408	0.18850397	0.01894595	0.07251050	0.008415829	0.184561578	1.00000000	
-0.05415801	pricexsqft_living	0.55452985	-0.09013088	-0.03124128	-0.08839251	-0.02240592	-0.29121236	0.101235020	0.039249119	-0.054158014
1.00000000										

Las correlaciones no parecen muy altas entre variables que había elegido inicialmente. Sin embargo no voy a incluir sqft_above y sqft_basement porque

VIF para poder analizar las variables categóricas:

	GVIF	DF	GVIF^(1/(2*DF))
sqft_living	3.204191	1	1.790025
grade	3.929686	11	1.064183
condition	1.315631	4	1.034884
waterfront	1.542789	1	1.242091
view	1.749462	4	1.072415
yr_built	1.797335	1	1.340647
sqft_living15	2.762081	1	1.661951
centro	1.210009	1	1.100004

VIF está bastante bajo, parece que no hay multicolinealidad

Modelo predictivo

Obtengo modelo de las siguientes características:

```
lm(formula = price ~ sqft_living + grade + condition + waterfront +
  view + yr_built + sqft_living15 + centro, data = Train)
```

Residuals:

Min	1Q	Median	3Q	Max
-1353395	-93630	-10210	81413	3880455

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.457e+06	2.423e+05	10.140	< 2e-16 ***
sqft_living	1.548e+02	3.165e+00	48.912	< 2e-16 ***
grade3	-1.126e+05	2.263e+05	-0.498	0.618796
grade4	-1.794e+05	2.037e+05	-0.881	0.378471
grade5	-2.009e+05	1.973e+05	-1.018	0.308529
grade6	-1.947e+05	1.971e+05	-0.987	0.323477
grade7	-1.609e+05	1.972e+05	-0.816	0.414508
grade8	-1.053e+05	1.972e+05	-0.534	0.593364

```

grade9      4.951e+03  1.973e+05  0.025  0.979985
grade10     1.778e+05  1.975e+05  0.900  0.368126
grade11     4.052e+05  1.979e+05  2.047  0.040628 *
grade12     9.553e+05  1.992e+05  4.796  1.63e-06 ***
grade13     1.811e+06  2.058e+05  8.801  < 2e-16 ***
condition2  9.784e+04  4.720e+04  2.073  0.038204 *
condition3  1.131e+05  4.363e+04  2.592  0.009539 **
condition4  1.286e+05  4.362e+04  2.949  0.003196 **
condition5  1.662e+05  4.389e+04  3.787  0.000153 ***
waterfront1 6.053e+05  2.305e+04  26.264 < 2e-16 ***
view1       1.108e+05  1.335e+04  8.302  < 2e-16 ***
view2       4.951e+04  8.224e+03  6.021  1.78e-09 ***
view3       1.160e+05  1.103e+04  10.523 < 2e-16 ***
view4       2.602e+05  1.693e+04  15.371 < 2e-16 ***
yr_builtin -1.236e+03  7.508e+01  -16.461 < 2e-16 ***
sqft_living15 4.108e+01  3.937e+00  10.432 < 2e-16 ***
centro      2.343e+05  3.835e+03  61.106 < 2e-16 ***

```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 192200 on 13667 degrees of freedom
Multiple R-squared: 0.7532, Adjusted R-squared: 0.7528
F-statistic: 1738 on 24 and 13667 DF, p-value: < 2.2e-16

Selección de variables

Todas las variables utilizadas parecen significativas, voy a hacer una comprobación adicional con backward selection.

```

stepAIC(modelolm,direction="backward") # veo con backward selection si excluir alguna variable
Start: AIC=333191.7
price ~ sqft_living + grade + condition + waterfront + view +
      yr_builtin + sqft_living15 + centro

      Df  Sum of Sq    RSS    AIC
<none>          5.0505e+14 333192
- condition      4 2.9226e+12 5.0798e+14 333263
- sqft_living15  1 4.0219e+12 5.0908e+14 333298
- yr_builtin     1 1.0013e+13 5.1507e+14 333458
- view           4 1.4299e+13 5.1935e+14 333566
- waterfront     1 2.5491e+13 5.3055e+14 333864
- sqft_living    1 8.8409e+13 5.9346e+14 335398
- grade          11 1.2657e+14 6.3163e+14 336232
- centro          1 1.3798e+14 6.4304e+14 336497

```

El mejor modelo sale sin excluir ninguna variable.

Validación del modelo

Pruebo el modelo en la muestra de test:

```

> R2_Train
[1] 0.7531961
> R2_Test
[1] 0.6869952

```

R cuadrado baja un poco, puede que tenga algo de overfitting, aunque la bajada de R^2 no parece considerable.

Comprobación con modelo robusto

Voy a comprobar qué resultado sale con el modelo robusto:

```

modelorlm
=rlm(price~sqft_living+grade+condition+waterfront+view+yr_builtin+sqft_living15+centro,data=Train)

> AIC(modelolm)
[1] 372049.9
> AIC(modelorlm) # el modelo robusto no sale mejor
[1] 373178.8

```

Modelos contraídos

Si hubiera observado multicolinealidad en mi modelo predictivo, probaría con modelos lasso o ridge. Sin embargo no la he observado, entonces me quedo con el modelo OLS.

Estimación de precios

Voy a estimar los precios de las viviendas en el data set house_test con ayuda de mi modelo predictivo. Para eso cargo los datos y formateo las variables categóricas. También añado la variable Dummy “centro”. A continuación aplico el modelo predictivo y estimo los precio de las casas nuevas.

El precio medio de las viviendas estimadas está alrededor de los 500 mil dolares y se parece al resto de las viviendas en Seattle:

```
price
Min. : 48927
1st Qu.: 329663
Median : 478192
Mean   : 546619
3rd Qu.: 671351
Max.   :4119913
```

Obtenemos los precios estimados con el siguiente perfil, bastante parecido al resto de la ciudad:

