Universidad del Valle de Guatemala
Minería de Datos - Sección 10
María Fernanda Estrada 14198
Christopher Kevin Sandoval 13660
20/03/2020

Hoja de Trabajo 4

Análisis del modelo

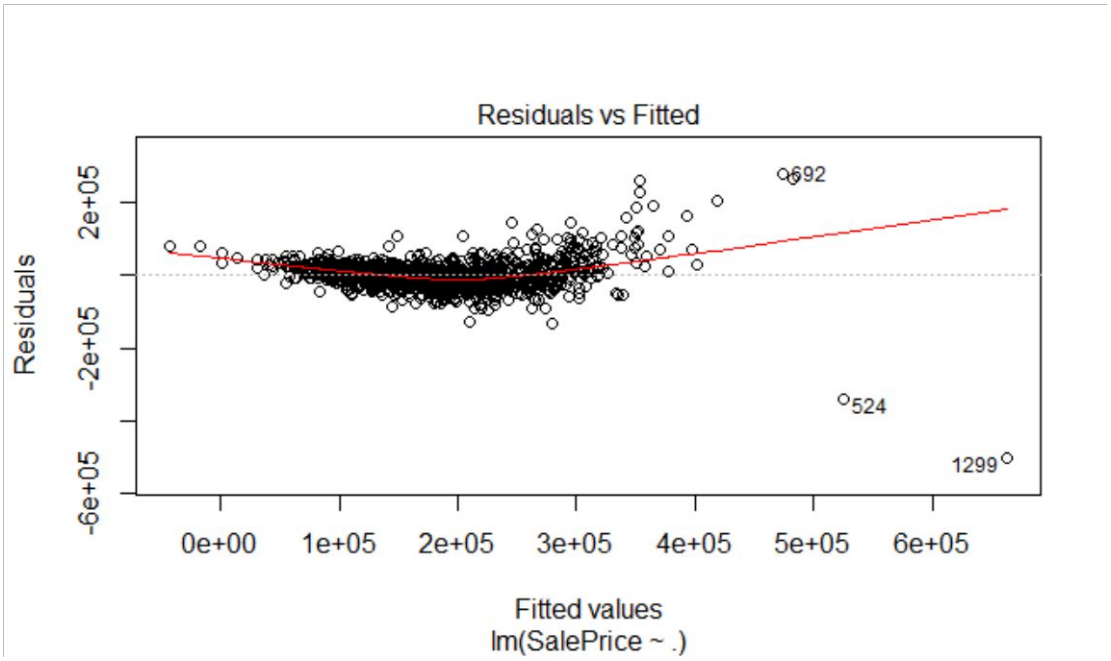
Se utilizaron las mismas variables que se consideraron importantes a la hora de predecir el precio de la casa. Al aplicar el modelo de regresión lineal al conjunto de datos, se obtuvo la siguiente información:

```
> summary(fitLMPW)

Call:
lm(formula = SalePrice ~ ., data = data_training_filtered)

Residuals:
    Min      1Q  Median      3Q     Max
-503494  -19959   -1737   16522  280107

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.126e+05  8.142e+04 -11.208   <2e-16 ***
OverallQual  2.265e+04  1.156e+03  19.591   <2e-16 ***
TotalBsmtSF  1.994e+01  4.352e+00   4.582    5e-06 ***
X1stFlrSF    1.819e+01  5.019e+00   3.624    3e-04 ***
GrLivArea    5.179e+01  2.737e+00  18.918   <2e-16 ***
YearBuilt    4.234e+02  4.282e+01   9.887   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38980 on 1454 degrees of freedom
Multiple R-squared:   0.76,     Adjusted R-squared:  0.7592
F-statistic:   921 on 5 and 1454 DF,  p-value: < 2.2e-16
```

Según esta información del modelo, el nivel de significancia de todas las variables es 0 y se tiene un R2 de 0.76, el cual es bastante aceptable. Esto quiere decir que todas las variables son importantes para predecir el precio de una casa.
Como la gráfica del modelo solo soporta dos variables, se fue probando una por una y observando que tuvieran una relación lineal con el precio de la casa. Algunos ejemplos son:

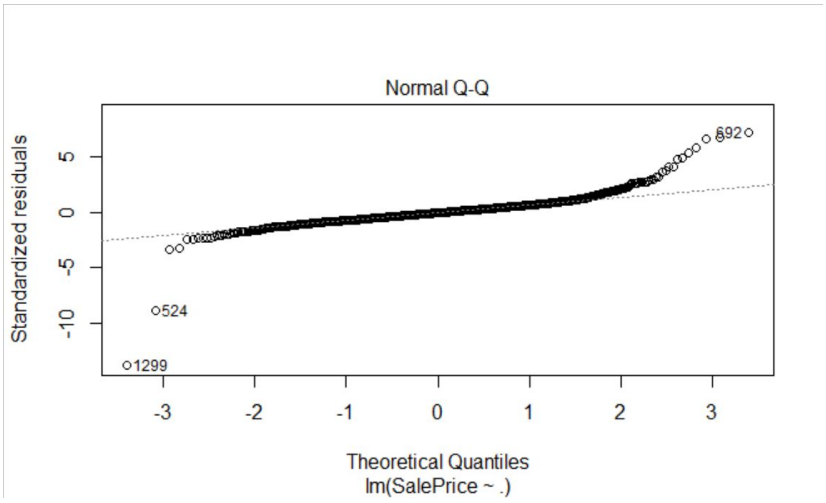
El análisis de residuos nos indica qué tan bueno será el modelo para predecir futuros datos. Se obtuvieron los siguientes resultados de residuos, indicando que el modelo predecirá bastante bien porque los datos son aleatorios:

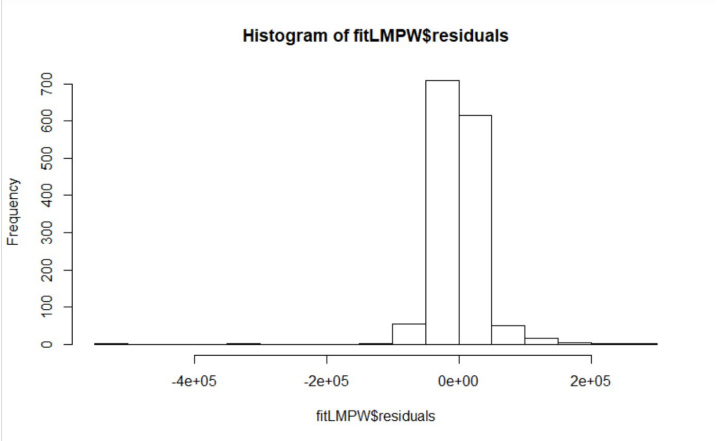
```
> head(fitLMPW$residuals)
         1          2          3          4          5          6
-6737.234  8063.244  2733.214 -38258.042 -22911.261  -2390.119
```



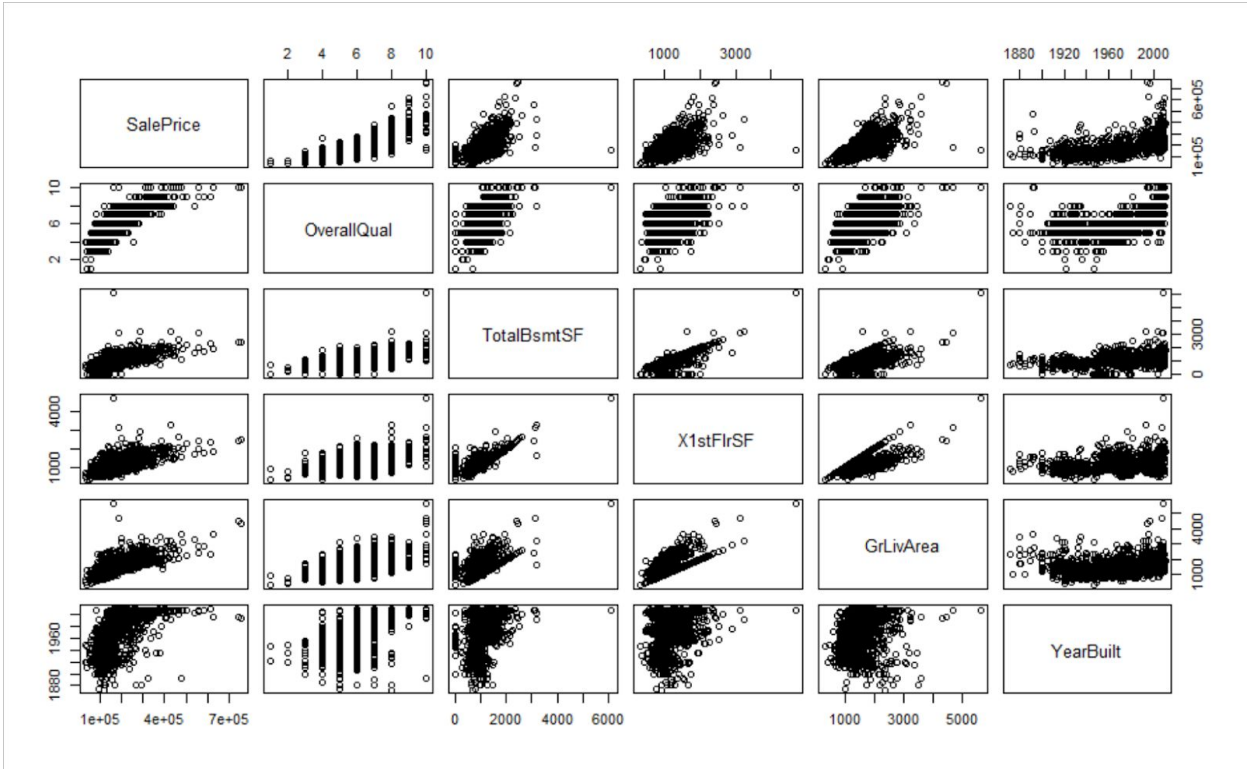
Análisis de las variables

Las pruebas de normalidad sobre los datos produjeron las siguientes gráficas:

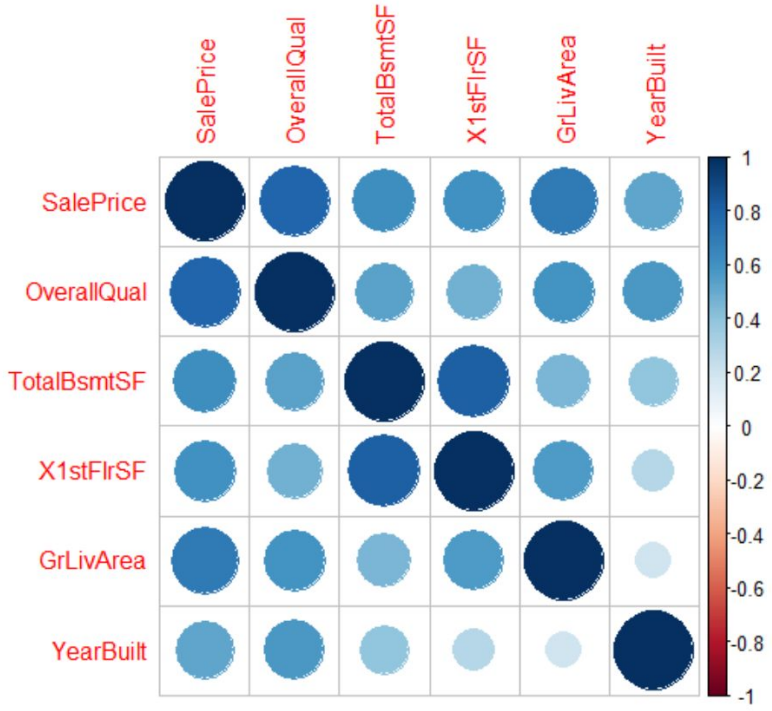




Ahora revisamos que no exista multicolinealidad entre las variables. Se observa en la gráfica siguiente que todas presentan relación lineal y que todas aportan al modelo.



Por último analizamos si existe correlación entre las variables del modelo. Se observa que hay entre 40% y 50% de correlación entre variables.



Aplicación del modelo

Par

Resultados obtenidos

Par

Comparación de métodos

Par