Regresión Lineal Múltiple
Airbnb en New York

Universidad Nacional Autónoma de México

Facultad de Ciencias

Modelos no paramétricos y de Regresión

Mayo 2020

Equipo:

Ávila Argüello Carlos

Bonilla Cruz José Armando

Luna Gutiérrez Yanely

Rivera Mata Dante Tristán

Introducción

Recabar datos es una herramienta muy útil para toda clase de áreas, puede ser para empresas, laboratorios, cuestiones políticas o bien poblacionales, pero no es importante sólo la recabación de estos. Lo que tiene peso es la forma en la que se interpretan. En nuestro caso decidimos analizar los datos relacionados a la información de AIRBnB en New York (NY) para ver si de alguna manera un conjunto de datos puede describir a un dato específico de forma lineal y así poder tener información de cómo es que son afectadas las variables.

Para contextualizar, AIRBnB es una empresa que ofrece una plataforma digital en la que se ofertan alojamientos a particulares y turísticos en la cual las personas pueden publicitar y arrendar sus propiedades a otras personas por un periodo de tiempo establecido.

Objetivo

En forma general lo que se busca de este trabajo, es poder construir un modelo de regresión lineal múltiple, aunque sea necesario darle tratamientos especiales a los datos como puede ser imputación de datos, análisis de outliers, pruebas acerca de sobre parametrización, transformación de variables, multicolinealidad, etc.

Análisis Descriptivo

Selección de variables

En este proyecto buscamos **explicar al Precio** de los Airbnb de la Ciudad de NY. Contamos con una base de datos de 16 variables:

- id (identificador único). Descartada pues no describe el precio.
- name (nombre del inmueble). Descartada pues no cambia el precio.
- host_id (llave única del arrendador). Descartada por no explicar el precio.
- host_name (nombre del arrendador). Descartada por falta de influencia en precio.
- **neighbourhood group** (distrito). Usada porque hay lugares más concurridos que otros y por tanto influyen en los precios. **Variable categórica**
- neighbourhood (barrio). Descartada, está contenida en el punto anterior.
- **latitude** (latitud). Usada por ser **continua** y porque influye en el precio.
- **longitude** (longitud). Usada por ser **continua**, complementa la anterior.
- **room type** (tipo de inmueble). Usada pues el precio depende del lugar de hospedaje. **Variable categórica.**
- **price** (precio). Usada pues es la variable dependiente del resto seleccionado. **Variable continua.**
- **minimum nights** (mínimo de noches). Usada porque creemos en su influencia en el precio. **Variable discreta.**
- number_of_reviews (número de reseñas). Descartada por no influir en el tiempo.
- last_review (última reseña). Descartada por no influenciar el precio.

- **reviews per month** (reseñas por mes). Considerada por ser **continua**, ponderar en el tiempo e influenciar a los clientes a rentar.
- calculated_host_listing_counts (número de propiedades del arrendador). Descartada, por no influenciar en el precio.
- **availability 365** (disponibilidad). Considerada por influenciar a los clientes a rentar. Es **variable discreta.**

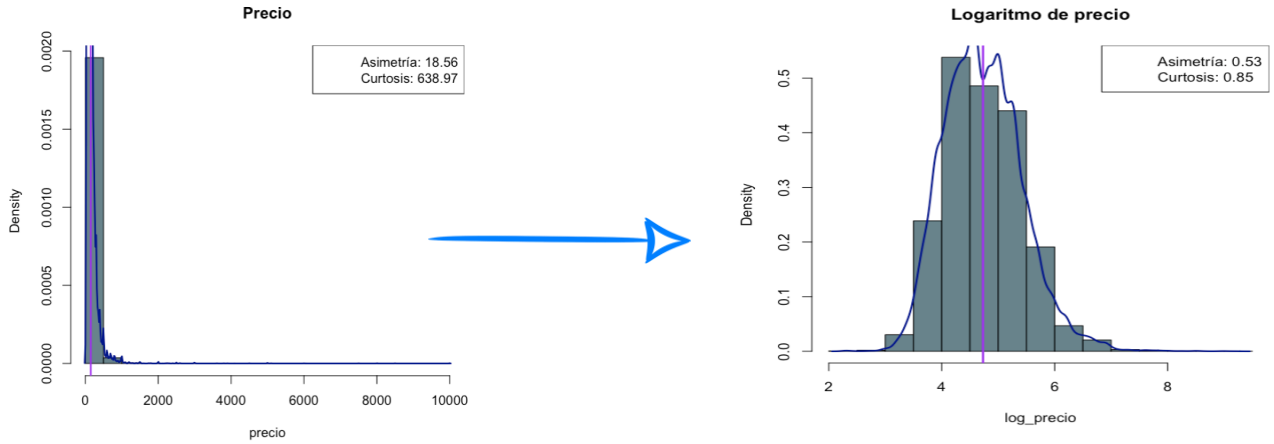
Filtros

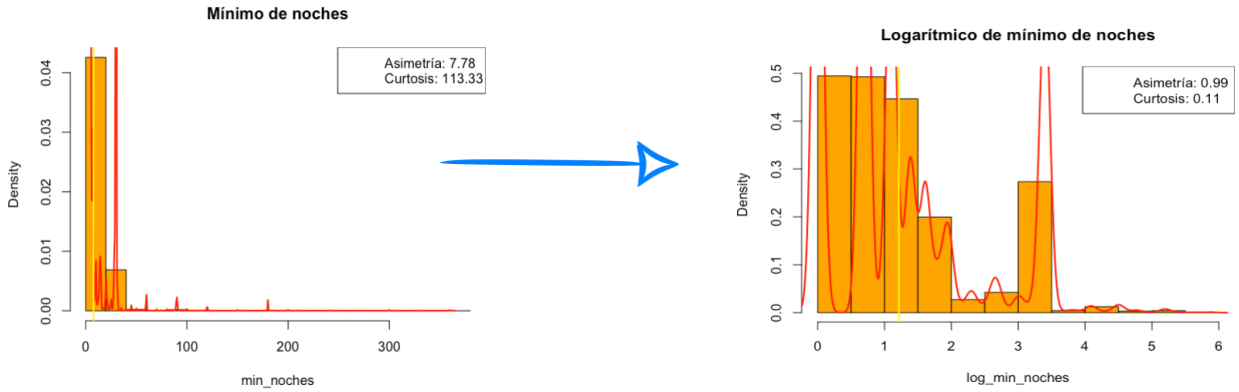
I. Precios positivos, los precios cero no podrían ser explicados con las variables seleccionadas.
II. Mínimo de noches en un rango de un día a un año. Dado el análisis de la fecha de forma mensual.
III. Disponibilidad desde un día, dado que con disponibilidad 0, no hay un precio.
IV. Debe haber al menos una crítica en la última fecha.

Modificación y transformación de variables

Dados los histogramas que se muestran a continuación, decidimos:

- **Recategorizar.** Dimos la categoría por mes para las fechas porque creemos que existen ciertos meses en donde hay más o menos demanda de lugares y esto afecta el precio.
- **Transformar la escala**. Dada la gran dispersión de los precios y mínimo de noches (ver histogramas) decidimos aplicar la escala logarítmica a ambas variables para tener un mejor manejo de datos.
- **Estandarización.** Para la latitud y longitud aplicamos una estandarización, es decir restamos su media y dividimos esta resta entre su desviación estándar. Esto con el objetivo de tener una fácil interpretación de los coeficientes de estas variables regresoras.





Variables explicativas. Distrito, latitud estándar, longitud estándar, tipo de inmueble, logaritmo del mínimo de noches, crítica por mes, fecha de última crítica y disponibilidad.
Variable respuesta. Logaritmo del precio.

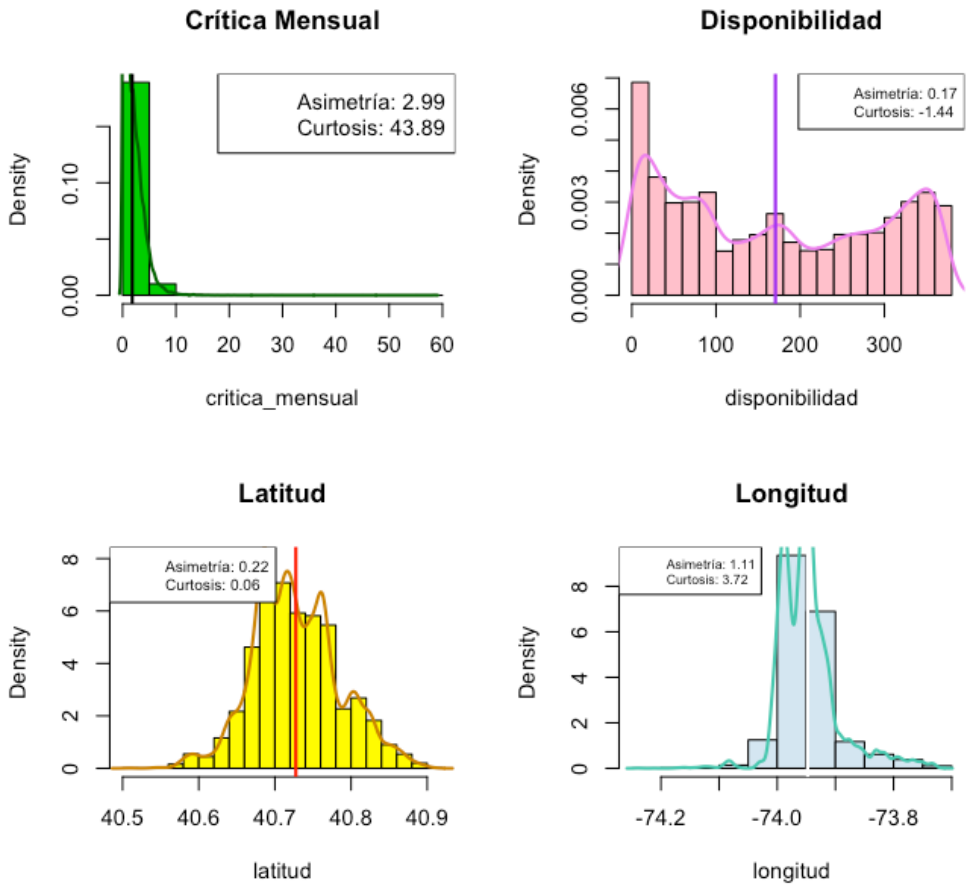
A partir de estas modificaciones en la base de datos, obtuvimos lo siguiente:

En la transformación que hicimos sobre la variable **precio** tenemos que el rango es de 2.303 a 9.210, lo que nos permite visualizar mejor la distribución de los datos en el histograma y en este notamos que la distribución tiene un sesgo a la izquierda y es platicúrtica.

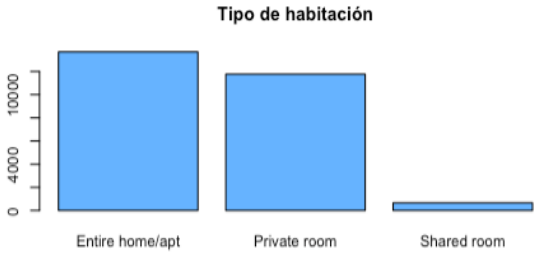
```
> summary(log_precio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.303   4.248   4.691   4.732   5.165   9.210
> summary(log_min_noches)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.6931  1.0898  1.3863  5.8972
```

En la transformación sobre el **mínimo de noches** tenemos ahora un rango de 0 a 5.89 y en el histograma de esta variable notamos que los primeros valores son los más altos, lo que nos dice que el mínimo de noches suele ser pequeño, pero aun así tenemos valores bastante altos que se aprecian en la cola de la distribución.

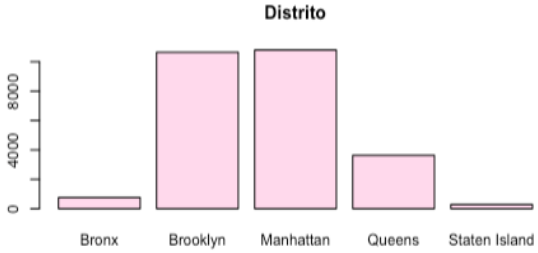
HISTOGRAMAS



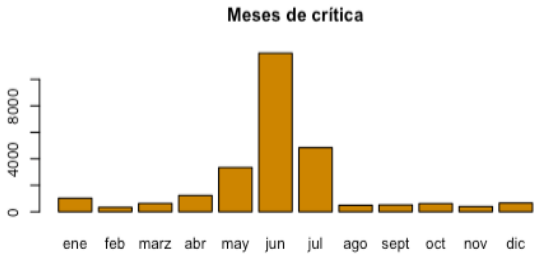
La **latitud** y la **longitud** se mueven entre rangos "pequeños", y eso está bien, pues sólo estamos recopilando la información de los Airbnb de la ciudad de NY. Por otra parte, el **número de críticas mensuales** tiene mayor frecuencia en "1" y "2" por la media, mediana y tercer cuartil, por tanto, el valor máximo es atípico (lo cual se confirma con el histograma), remarcando el sesgo a la derecha de esta distribución. La **disponibilidad** está mejor distribuida con media aproximada a medio año.



Casi la mitad de nuestras observaciones de **tipo de habitación** son casa/departamentos enteros y una proporción muy pequeña son cuartos compartidos.

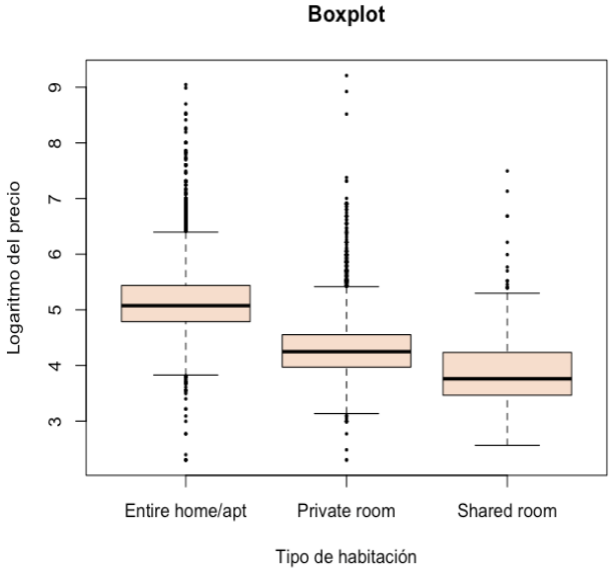


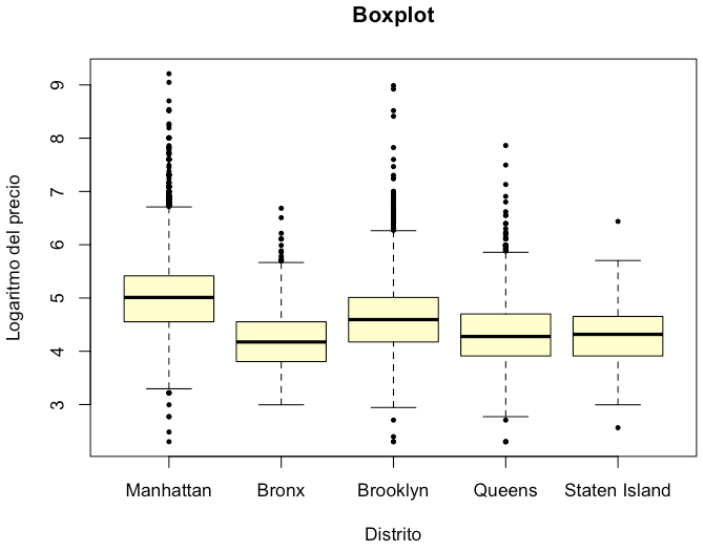
Los **distritos** con más solicitudes son Brooklyn y Manhattan, justamente los que están más cerca de las zonas turísticas.



Podemos observar de la variable **Mes de crítica**, que hay muchas rentas de Airbnb para los meses de mayo, junio y julio, esto tiene sentido pues son los meses vacacionales del año.

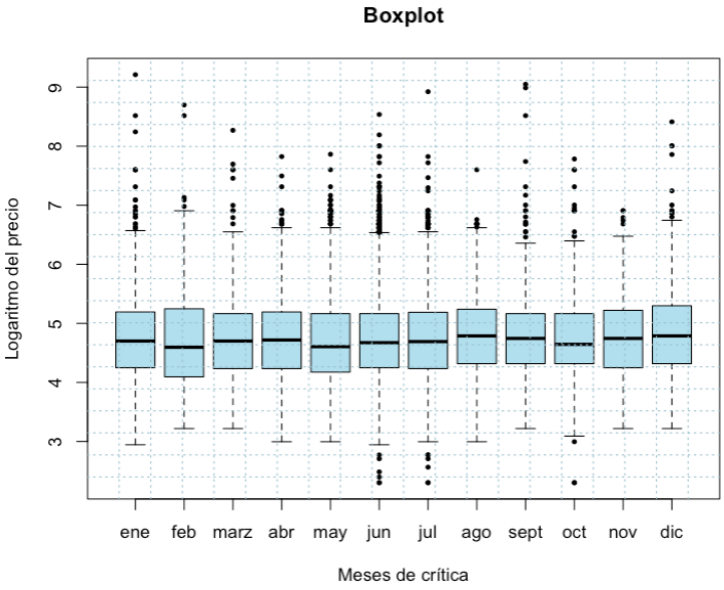
En cada una las gráficas de boxplot podemos ver que las rectas de cuantiles .5, difieren mucho entre todas, esto nos indica que si existe una diferencia significativa del Log(precio) según el tipo de habitación que se esté rentando, por lo tanto, es adecuado dejar todas las categorías para el análisis de regresión lineal múltiple.





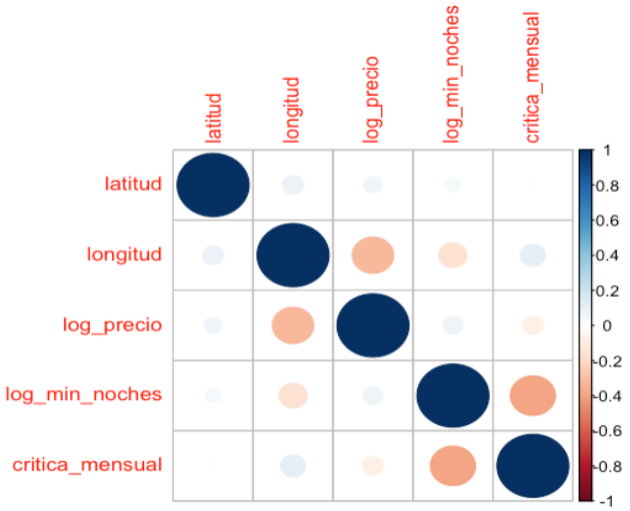
En este boxplot conjunto es donde podemos apreciar que el logaritmo de los precios sí depende de del distrito, por ejemplo, en Manhattan y en Brooklyn es donde tenemos más valores atípicos. y en promedio, los Airb&b de Manhattan son los más caros de la ciudad

El boxplot de la variable del mes donde se realizó la última crítica nos dice que en todos los meses se observa una mediana muy similar del logaritmo del precio, al igual que para el primer y tercer cuantil. En el mes de junio y julio se observan varios valores atípicos muy pequeños y varios más que son muy altos, esta variabilidad puede estar causada porque son los meses vacacionales y son los que tienen más observaciones. El mes de octubre solo tiene un par de valores atípicos pequeños. El resto de los meses presenta valores atípicos altos en distinta medida, siendo noviembre el que tiene menor cantidad de estos.



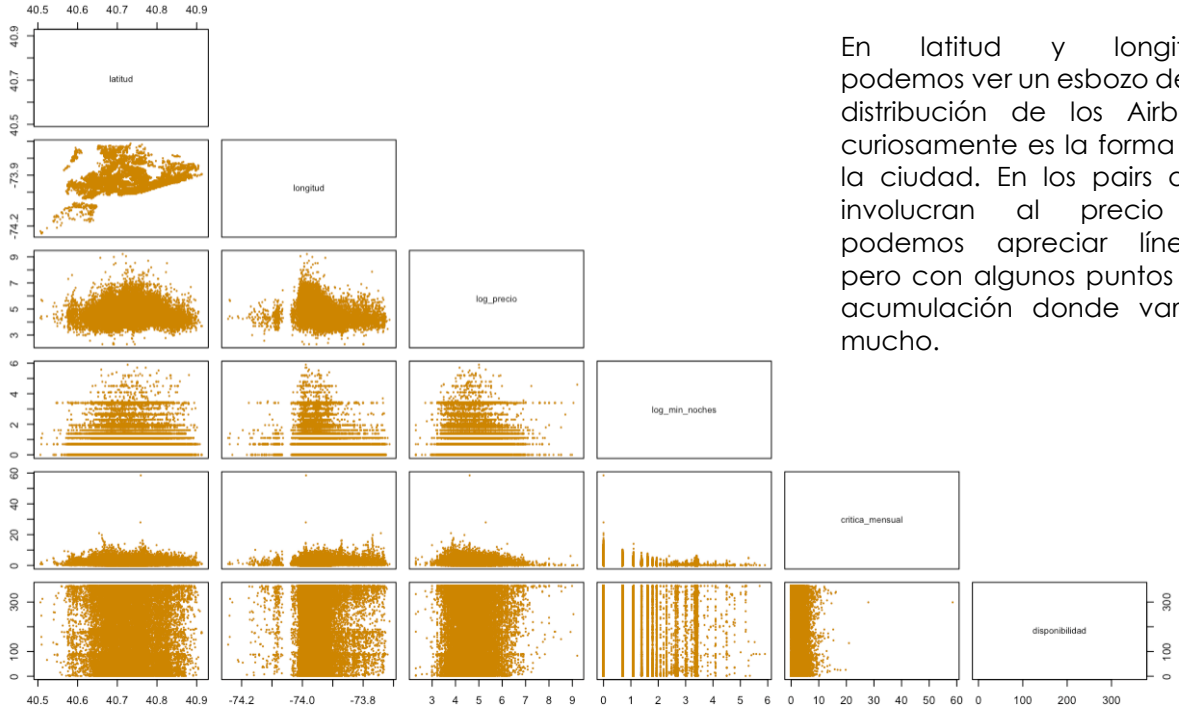
EN CONJUNTO

Correlaciones



En este esquema de correlación, solo tomamos en cuenta las variables son continuas para que tenga sentido hacer dicho esquema, además vemos que se tiene una correlación casi nula para la mayoría de las combinaciones de variables, a excepción de longitud-log precio y critica mensual con log_min_noches.

Pares de variables



En latitud y longitud podemos ver un esbozo de la distribución de los Airbnb, curiosamente es la forma de la ciudad. En los pairs que involucran al precio sí podemos apreciar líneas, pero con algunos puntos de acumulación donde varían mucho.

Análisis estadístico

Outliers

Para obtener los "Outliers" consideramos el criterio de las notas de clase, el cual nos dice que, a un nivel de significancia α, tenemos que si

$$|r_i| \geq t_{1-\alpha/2,n-p}$$

entonces, la i-ésima observación de la muestra es un posible outlier.

Donde r_i es el error estudentizado de la i-ésima observación. Es decir que si el valor absoluto de este error estudentizado es mayor o igual al cuantil $1-\alpha$ de una v.a. que tiene distribución t de student con $n-p$ grados de libertad, estamos en la posibilidad de encontrar un outlier.

Dado que no tenemos información sobre la recopilación de los datos, no podemos identificar que los outliers encontrados sean corregidos por una mala captura, por lo tanto ,son datos atípicos y los quitamos del modelo, para crear un modelo_sda = modelo sin datos atípicos.

CONSTRUCCIÓN DEL MODELO

Tabla comparativa de las transformaciones realizadas

Modelo	Transformación	Coeficiente de determinación ajustado	Supuestos del modelo					Outliers
			No correlación		Homocedasticidad		Normalidad	
			Brush-Godfry	Durbin-Watson	Brush-Pagan	Non Constant Variance	Anderson-Darlin	
Con atípicos								
Sin transformaciones		14.22%	1.72*10^-18	8.48*10^-19	1.38*10^-15	0	3.7*10^-24	1.74%
1	log(precio)	53.74%	2.05*10^-16	9.28*10^-17	9.07*10^-85	1.41*10^-119	3.7*10^-24	4.89%
2	(log(precio))^lambda	56.64%	3.55*10^-15	1.5*10^-15	1.55*10^-55	1.27*10^-22	3.7*10^-24	5.00%
3	(log(precio))^(-1/2)	56.60%	3.5*10^-15	1.46*10^-15	1.74*10^-56	1.07*10^-34	3.7*10^-24	4.98%
4	(log(precio))^-lambda	55.35%	6.05*10^-16	2.65*10^-16	8.27*10^-62	3.78*10^-20	3.7*10^-24	5.05%
5	(log(precio))^(1/2)	55.11%	4.97*10^-16	2.19*10^-16	2.16*10^-65	1.29*10^-31	3.7*10^-24	5.02%
Sin atípicos								
Sin transformaciones		50.28%	1.005*10^-7	4.93*10^-8	0	0	3.7*10^-24	4.74%
1	log(precio)	64.48%	2.56*10^-8	1.13*10^-8	3.75*10^-68	2.22*10^-43	3.7*10^-24	5.28%
2	(log(precio))^lambda	65.09%	8.94*10^-9	3.79*10^-9	1.36*10^-29	0.00015	3.7*10^-24	5.01%
3	(log(precio))^(1/3)	65.04%	1.04*10^-8	4.46*10^-9	2.33*10^-26	0.158	3.7*10^-24	4.99%
4	log(log(precio))	65.17%	6.11*10^-9	2.56*10^-9	1.6*10^-46	1.99*10^-20	1.42*10^-20	4.87%
5	sqrt(log(precio))	64.96%	1.33*10^-8	5.76*10^-9	9.33*10^-27	0.013	3.7*10^-24	5.11%
Ideal	?	65.17%	1.005*10^-7	4.93*10^-8	1.38*10^-15	0.158	1.42*10^-20	1.74%

Cada renglón tiene también la transformación log(min_noches) excepto los renglones "Sin transformaciones". log=logarítmo natural. lambda = el exponente después de aplicar el método de "BoxCox"

Para poder decidir cuál era el mejor modelo de regresión lineal múltiple construimos una función en R, la cual pedía las variables cuantitativas (explicativas) y la variable respuesta con sus respectivas transformaciones y dejaba fijas las variables categóricas como factores (pues estas no se pueden transformar). Lo que nos arroja esta función son las pruebas paramétricas acerca de los supuestos que tiene que cumplir para poder ser considerado un modelo de regresión lineal múltiple. Las pruebas paramétricas realizadas, son las que se

encuentran en la sección de "Supuestos del modelo" en la tabla anterior. Dentro de la tabla se colocaron los p-values de cada prueba, así como la R^2_{adj} y el porcentaje de outliers (que corresponde a los residuos de outliers, respecto a los datos de la transformación). Obsérvese que en la tabla se separan transformaciones considerando: todas observaciones o sin considerar los datos atípicos (quitando los outliers como se explica en la sección anterior).

En la tabla se puede observar que los p-values obtenidos son pequeños para rechazar la hipótesis nula de su respectiva prueba excepto uno (celda color verde). Como en casi todas las celdas el valor del p-value rechaza la prueba nos indica que en ninguna de las transformaciones realizadas funciona para cumplir los supuestos en cuanto a las pruebas de hipótesis.

A pesar de rechazar los supuestos con todas las transformaciones, **consideraremos** como modelo de regresión lineal múltiple -de este proyecto- **el modelo que elimina los datos atípicos y tiene como transformaciones: el logaritmo del precio, el logaritmo del mínimo de noches, la estandarización de la latitud y longitud además del resto de variables explicativas sin transformar.**

Por lo tanto, el modelo para el precio de los AirBnB en la ciudad de New York es

$$y = \beta_0 + \sum_{i=1}^{22} \beta_i x_i$$

Donde:

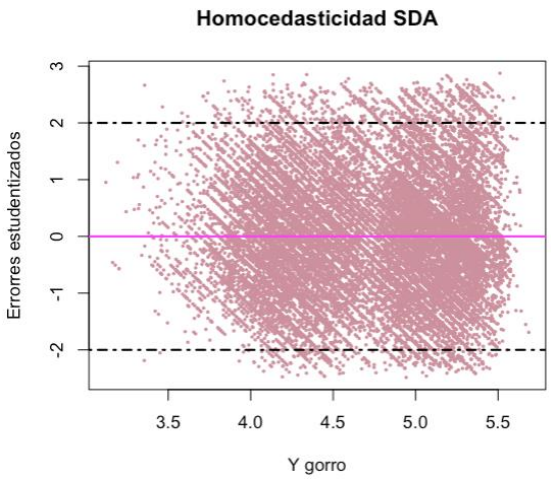
- i son los pesos de las variables explicativas
- x_i son los valores asociados a cada observación, por ejemplo x_1 es el log mínimo de noches de la observación.

El comportamiento de nuestras variables dummies será un poco diferente, si por ejemplo queremos calcular el log precio de un Airbnb que tiene la subcategoría "Shared Room" de la categoría "Tipo de Habitación", entonces la variable x_1 asociada a la subcategoría "Shared Room" tendrá el valor de 1, y todas las demás variables de la misma categoría tendrán el valor de 0. Podemos notar que son muchos parámetros, dado que tenemos variables categóricas, y por cada categoría tenemos (#sub-catergorias - 1) parámetros.
La variable y representa el log_precio, como queremos calcular el precio, tenemos que:

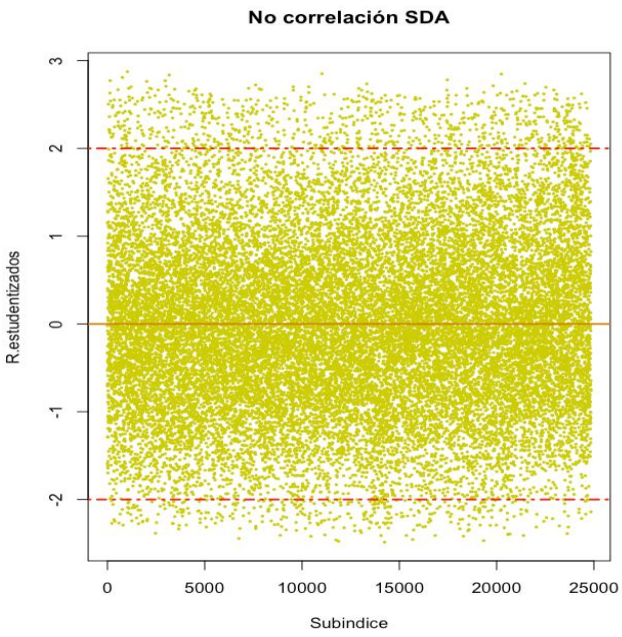
$$Precio = exp(y)$$

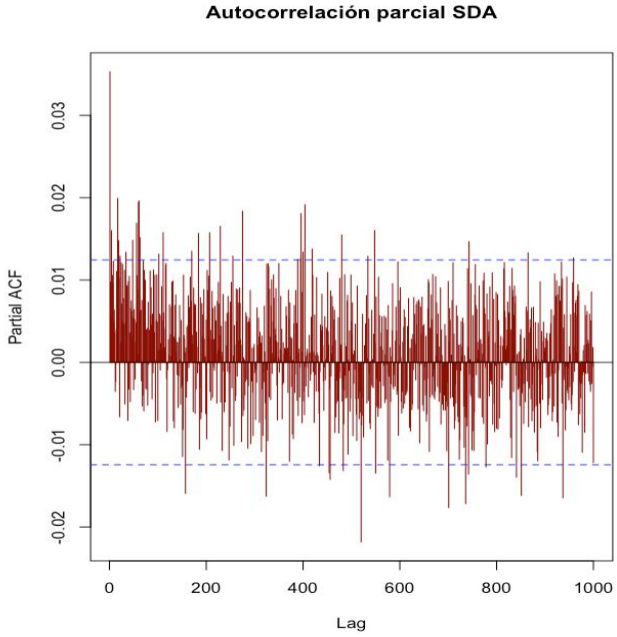
La razón por la cual decidimos tomar el modelo anterior es que gráficamente se comprueban los supuestos, aunque esto no se refleja en las pruebas estadísticas. Esto tiene que ver con el hecho de que las pruebas de hipótesis dependen del tamaño de muestra, y dado que nuestra muestra ronda entre los 26 mil datos, es de esperarse que exista una influencia fuerte en las pruebas para rechazar H_0. A continuación se presentan algunas gráficas que comprueban los supuestos con las transformaciones elegidas:

En la **gráfica** de la **derecha** graficamos los valores ajustados contra los residuos estudentizados. Podemos ver que la mayoría de los datos/puntos se limitan a las rectas y que dichos puntos no tienen una dispersión aleatoria, lo que nos indicaría que sí tenemos **varianza constante**.

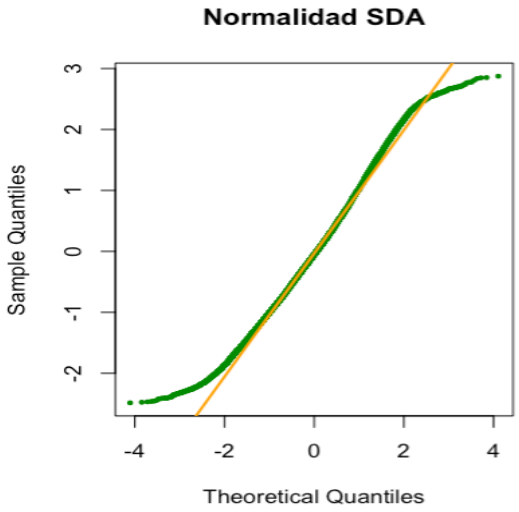


En la gráfica de los residuos estudentizados contra el subíndice no se observa ningún patrón sobre estos datos, lo cual nos dice que el residuo de una observación no se ve influenciado por el del anterior y todos se encuentran en el rango de -3 a 3 dispersos de forma aleatoria. Esto es una señal de que los **residuos no** están **correlacionados**. La gráfica siguiente, de autocorrelación parcial, apoya esta afirmación pues las autocorrelaciones con "lags" entre 1 y 1000 tienen la mayoría valores absolutos menores a 0.01.





Para evaluar el supuesto de **normalidad**, podemos observar en la gráfica de QQ plot que los cuantiles del modelo se parecen mucho a los cuantiles teóricos en el "centro de la distribución", pero si nos fijamos bien, las colas de la distribución son más "pesadas" que las teóricas. Para verificar nuevamente la normalidad de los errores graficamos el histograma y nos percatamos que la forma es factible, así como que su media se acerca mucho a 0.



Resumen del modelo

```
Call:
lm(formula = log_precio ~ log_min_noches + latitud_est + longitud_est +
    critica_mensual + disponibilidad + distrito + habitacion +

Residual standard error: 0.362 on 24823 degrees of freedom
Multiple R-squared:  0.6452,    Adjusted R-squared:  0.6449
F-statistic:  2052 on 22 and 24823 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)               5.297e+00  8.395e-03  631.027  < 2e-16 ***
log_min_noches           -9.524e-02  2.568e-03  -37.087  < 2e-16 ***
latitud_est              -1.970e-02  3.854e-03   -5.113 3.19e-07 ***
longitud_est             -1.434e-01  3.787e-03  -37.874  < 2e-16 ***
critica_mensual          -2.051e-02  1.601e-03  -12.812  < 2e-16 ***
disponibilidad            4.389e-04  1.888e-05   23.250  < 2e-16 ***
distritoBronx            -3.489e-01  1.596e-02  -21.861  < 2e-16 ***
distritoBrooklyn         -2.789e-01  8.114e-03  -34.372  < 2e-16 ***
distritoQueens           -1.869e-01  1.118e-02  -16.712  < 2e-16 ***
distritoStaten Island    -1.055e+00  2.574e-02  -40.976  < 2e-16 ***
habitacionPrivate room   -7.860e-01  4.899e-03 -160.449  < 2e-16 ***
habitacionShared room    -1.316e+00  1.496e-02  -87.931  < 2e-16 ***
mesesene                  1.695e-02  1.243e-02    1.364  0.17270
mesesfeb                 -6.795e-02  2.111e-02   -3.218  0.00129 **
mesesmarz                -2.216e-02  1.548e-02   -1.431  0.15243
mesesabr                 -7.774e-03  1.148e-02   -0.677  0.49819
mesesmay                 -1.163e-03  7.558e-03   -0.154  0.87768
mesesjul                  8.335e-03  6.418e-03    1.299  0.19404
mesesago                  3.183e-02  1.748e-02    1.821  0.06867 .
mesessept                 5.950e-02  1.717e-02    3.466  0.00053 ***
mesesoct                  1.028e-01  1.577e-02    6.518 7.25e-11 ***
mesesnov                  1.668e-02  1.927e-02    0.866  0.38674
mesesdic                  4.453e-02  1.543e-02    2.886  0.00390 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En el resumen podemos ver que más de la mitad de las variables explicativas se encuentran relacionadas inversamente con el logaritmo del precio.

Cabe mencionar que para todas las variables explicativas se obtuvieron VIF menores a 10, por lo que no asumimos problemas de multicolinealidad, además de comprobar de que la diferencia de la R^2 y la R^2_{adj} no es significativa.

	GVIF	Df	GVIF^(1/(2*Df))
log_min_noches_sda	1.346871	1	1.160548
latitud_est_sda	2.821214	1	1.679647
longitud_est_sda	2.700489	1	1.643317
critica_mensual_sda	1.518007	1	1.232074
disponibilidad_sda	1.037636	1	1.018644
distrito_sda	7.350687	4	1.283190
habitacion_sda	1.118159	2	1.028314
meses_sda	1.461111	11	1.017386

Análisis de varianza (ANOVA)

En la siguiente imagen podemos ver la tabla ANOVA, donde tenemos un resumen de la significancia de nuestras variables explicativas, es por eso que no se encuentra el intercepto, otra cosa que podemos observar es que nuestras variables categóricas no están desglosadas por sub-categorías.
Tenemos la siguiente prueba de hipótesis.

$$Ho: \beta j = 0 \quad vs \quad H1: \beta j \neq 0, \forall i$$

Esta prueba la vamos a hacer por cada j, donde cada j es una variable explicativa
Como podemos recordar, para esta prueba necesitamos la siguiente estadística,

$$F_j = MeanSq_j / MeanSq_{residuals}$$

lo **que significa el cociente entre la suma de los cuadrados de la regresión medios reducidos entre la suma de los cuadrados medios de los erroes.** que se distribuyen $F_{(k,24823)}$, donde K es 1 si nuestra variable j es cuantitativa, en otro caso, si nuestra variable es cualitativa k=#categorias-1, dada nuestra variable cualitativa j, rechazamos H0 con un nivel de significancia si $F_j > F^{1-\alpha}_{(k,24823)}$.
En la cuarta columna podemos ver el valor de la Fj, y en la quinta, el p-value por cada j, donde podemos observar que todas nuestras variables son significativas.

```
Analysis of Variance Table

Response: log_precio
                  Df Sum Sq Mean Sq    F value    Pr(>F)
log_min_noches     1   60.6   60.58   462.2611 < 2.2e-16 ***
latitud_est        1   34.8   34.77   265.3159 < 2.2e-16 ***
longitud_est       1 1180.1 1180.13  9004.5136 < 2.2e-16 ***
critica_mensual    1    1.2    1.24     9.4568  0.002106 **
disponibilidad     1   12.6   12.62    96.2762 < 2.2e-16 ***
distrito           4  733.0  183.25  1398.2081 < 2.2e-16 ***
habitacion         2 3882.4 1941.20 14811.5490 < 2.2e-16 ***
meses             11   10.9    0.99     7.5751 3.535e-13 ***
Residuals      24823 3253.3    0.13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Intervalos de confianza de los parámetros.

Calculamos los intervalos al 95% de confianza para los 23 parámetros de nuestro modelo (son 23 porque dentro de las categorías de meses, tipo de habitación y distrito, existen las subcategorías). Podemos notar que el parámetro de intercepto es el que se encuentra más alejado del cero y tiene un intervalo de confianza contenido en los positivos junto con otros 4 parámetros (disponibilidad, y 3 de la categoría meses: septiembre, octubre y diciembre),

por lo que aumentar la variable asociada a estos parámetros resulta en un aumento de la variable respuesta (log_precio).

También tenemos 11 parámetros (latitud_est, longitud_est, los de tipo de habitación, los de la categoría distrito y el mes de febrero) con intervalos contenidos en los negativos en los que el aumento (o la presencia en el caso de las categóricas) de las variables asociadas a estos parámetros se relaciona con una disminución del log_precio.

Por último tenemos 7 parámetros con intervalos que contienen al cero, que son los que tienen "menos significancia" en el modelo, tales son subcategorías de la categoría "meses" que tiene por categoría base "junio", lo cual no necesariamente quiere decir que no sean significativos en el cambio del log_precio **pero** quiere decir que el cambio del log_precio promedio para alguno de esos 7 meses, es parecido al cambio que se genera para el log_precio del mes de junio.

```
> confint(modelo_sda)
                                2.5 %       97.5 %
(Intercept)              5.2807796072  5.3136875382
log_min_noches          -0.1002762082 -0.0902090399
latitud_est             -0.0272580342 -0.0121513141
longitud_est            -0.1508574998 -0.1360115995
critica_mensual         -0.0236457268 -0.0173708075
disponibilidad           0.0004019033  0.0004759042
distritoBronx           -0.3802041450 -0.3176358300
distritoBrooklyn        -0.2948113550 -0.2630023217
distritoQueens          -0.2088436320 -0.1649978096
distritoStaten Island   -1.1049574284 -1.0040729493
habitacionPrivate room  -0.7956363962 -0.7764318628
habitacionShared room   -1.3451973940 -1.2865339507
mesesene                -0.0074131893  0.0413094489
mesesfeb                -0.1093341171 -0.0265634523
mesesmarz               -0.0525010852  0.0081899654
mesesabr                -0.0302693108  0.0147214410
mesesmay                -0.0159784302  0.0136516927
mesesjul                -0.0042441298  0.0209139714
mesesago                -0.0024363018  0.0660874036
mesessept                0.0258466184  0.0931451549
mesesoct                 0.0719022981  0.1337402934
mesesnov                -0.0210945729  0.0544588563
mesesdic                 0.0142874830  0.0747637784
```

Aplicación

Ejemplo 1

Una persona decide contratar un servicio de Airbnb en New York para pasar las vacaciones de verano (el mes de junio se tuvo como última crítica, 1 semana) con su familia, y quiere estimar el precio de este viaje, por lo que requiere saber el costo de esta contratación. Sus hijas quieren ir a Manhattan (Latitud:40.777594, longitud: -73.962734) pero todos han decidido quedarse en un departamento. Tal departamento tiene una disponibilidad de 365 días y una crítica mensual de dos.

El **precio por noche** obtenido bajo el modelo (calculado en R) es: $**191.63**, que se encuentra en el intervalo de predicción (94.24421, 389.66458) con una significancia del 5% y en el intervalo de confianza (189.2204, 194.0786).

Ejemplo 2

Una pareja decide arrendar para Airbnb en New York para que quien guste pueda pasar como mínimo 3 días del mes de agosto, y quiere estimar el precio de su renta por lo que requiere saber el costo de contratación. La pareja tiene en Queens (Latitud: 40.746922, longitud: -73.834151) una recamara compartida con una disponibilidad de 150 días y una crítica mensual de 0.1.

El **precio por noche** obtenido bajo el modelo (calculado en R) es: $**31.89186**, que se encuentra en el intervalo de predicción (15.66348, 64.93388) con significancia del 5% y en el intervalo de confianza (30.48459, 33.36408).

De los resultados podemos observar que es lógico que el precio obtenido en el ejemplo 1 sea mayor al del ejemplo 2 porque las características del apartamento del primer ejemplo son las que, según nuestro modelo, están asociadas con un mayor precio. Además, en ambos casos el intervalo de predicción es de mayor longitud que el de confianza, lo cual concuerda con la teoría de regresión lineal.

Los intervalos le permiten a los arrendatarios o arrendadores estimar un monto máximo y mínimo que deben esperar que se les cobre o al que puedan arrendar.

Conclusiones

En este trabajo nos enfrentamos a rechazar los supuestos de modelo por parte de las pruebas estadísticas, pero a su vez encontramos que los supuestos se cumplen gráficamente. Con lo anterior nos dimos cuenta de que la cantidad de datos que se tienen pueden afectar de manera significativa las pruebas de hipótesis.

A fin de cuentas, decidimos que nuestros datos cumplían los supuestos y ajustamos el modelo teniendo en cuenta las transformaciones realizadas, porque como se explicó, podemos interpretar el logaritmo del precio como un escalamiento del precio para no tener valores tan grandes.

Una forma de obtener un modelo más preciso podría ser incluyendo otro tipo de variables que pueden influenciar de manera importante en el precio, tales como los metros cuadrados de la habitación que se renta y el puntaje obtenido en las críticas que se reciben, no solo el número de éstas.