Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias

Instituto Tecnológico de Estudios Superiores y de Monterrey

Campus Puebla

Análisis de Datos y Herramientas de inteligencia Artificial II

Grupo 502

Profesor:

Alfredo García Suárez

Actividad:

Actividad 1 - Reporte comparativo Regresión Lineal

Alumno:

Jose Angel Fernandez Perez A01734131

23 de Octubre de 2022

Descripción

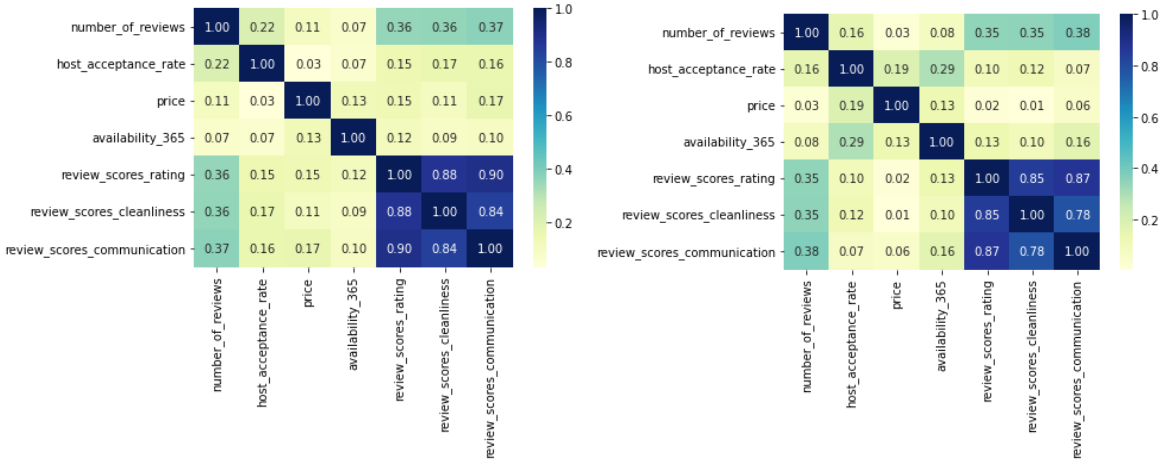
Con la base de datos de Airbnb de Ciudad de México se hizo un análisis de correlación de las variables dadas con el número de reviews. Tras la limpieza de la base de datos se hizo un modelo de regresión lineal de la variable independiente 'review_scores_communication' con la variable dependiente 'number_of_reviews' para los tipos de habitación *Entire Home/apt, Private Room* y *Shared Room*

Para los modelos de regresión lineal múltiple se eligieron las ciudades de Toronto y München. De igual manera se hizo la limpieza de datos y se encontró que las variables 'review_scores_rating', 'review_scores_cleanliness' y 'review scores communication' eran las de mayor correlación con el número de reviews, por lo que se eligieron dichas variables como variables predictoras.

Resultados

Tanto para la Ciudad de México como para Toronto se encontraron correlaciones muy similares de las variables independientes con respecto al número de reviews como se puede observar en la figura 1. Por lo cual, son las mejores para comprobar si la relación lineal simple y múltiple tuvo una mejora significativa en el modelo de predicción

Figura 1



Nota: Mapa de correlaciones a) Correlaciones de Ciudad de México, b) Correlaciones de Toronto

Para la ciudad de México se eligió la variable *review scores communication* con los resultados mostrados en la figura 2, Mientras que para Toronto se hizo un modelo de regresión lineal múltiple con las tres variables de *review* cuyos resultados igualmente se muestran en la figura 2. Se puede observar que hay una mejora en la correlación y la R cuadrada, pero es muy baja, con lo cual, agregar estas variables fue poco útil, probablemente el número de review dependa de otras variables extra.

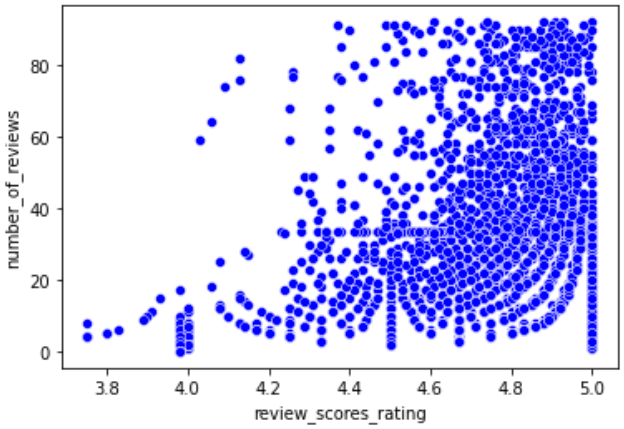
Agregar el precio o la disponibilidad no era factible, ya que tenían una correlación demasiado pequeña. y tendrán poco efecto en el modelo de regresión múltiple.

Figura 2

Tipo cuarto	Coef determ. Reg. Simple	Correlación Reg. Simple	Coef determ. Reg. Múltiple	Correlación Reg. Múltiple
Entire Home/apt	0.111	0.333	0.152	0.389
Private Room	0.139	0.373	0.151	0.388
Shared Room	0.121	0.347	0.121	0.348

Graficando una de nuestras variables *review scores rating* con respecto a nuestra variable de respuesta (figura 3), podemos observar que existe una tendencia: conforme crece una variable, crece la otra; pero en valores altos de la variable independiente, hay mucha dispersión de los datos. Con lo cual, a valores altos, hay otra variable que influye en el número de *reviews*

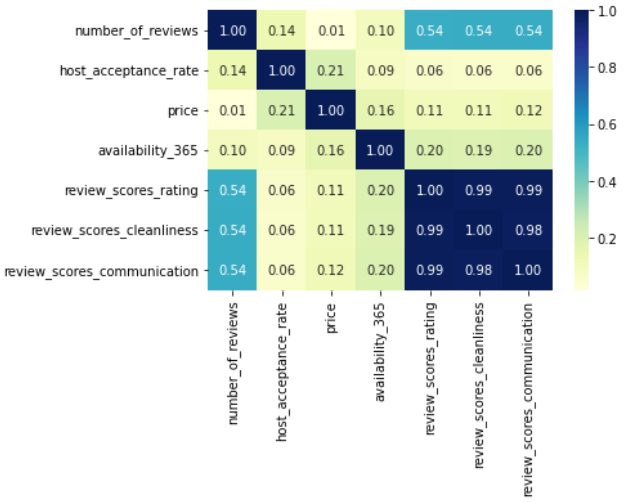
Figura 3



Nota: Gráfica de dispersión de *review scores rating* con *number of reviews*

Para el caso de München, se encontró una mejor correlación con todas las variables, como se observa en la figura 4. En el cual las 3 variables elegidas (las variables de *review)* tiene correlación de 0.54 cada una.

Figura 4



Nota: Mapa de correlaciones de München

En el caso de München el modelo de regresión lineal múltiple tuvo un coeficiente de correlación de 0.58 en promedio para los 3 tipos de habitaciones, con lo cual se concluye que hubo un aumento con respecto a la variables individuales, pero que este fue pequeño.

Conclusiones

Haber utilizado las variables con mayor correlación en el modelo de regresión lineal múltiple tuvo cierta mejora en la correlación comparado con la regresión simple. Pero esta fue muy pequeña, y esto se puede deber a que hay una muy alta correlación entre los tipos de review, lo que está generando información redundante. Aún así agregar más variables tuvo una mejora en el modelo, aunque faltarían otras variables que puedan explicar la variabilidad cuando hay un alto número de *reviews*