Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Puebla

Analítica de datos y herramientas de inteligencia artificial I
TI3001C.102

ACTIVIDAD 1. Regresión Lineal | Módulo Estadística

Por:
María Teresa Hernández Cid A01734400

22/10/2022

Reporte: Regresión Lineal Simple y Múltiple.

Introducción.

El siguiente reporte tiene el objetivo de comparar la dependencia del número de reseñas con respecto a demás variables de un conjunto de datos de Airbnb de 3 ciudades diferentes, siendo la ciudad de México en México, Berlín en Canadá y Hovedstaden, Dinamarca. Con el objetivo de corroborar correlaciones para comprobar dependencias, con el fin de realizar el mejor modelo de regresión lineal simple o múltiple que describiera mediante predicciones a nuestro atributo dependiente de 'número de reseñas'.

Metodología.

Antes de realizar nuestro objetivo de los modelos de regresión que mejor describieran al conjunto de datos. Se siguió la siguiente metodología para las 3 Ciudades.

- Extraer el conjunto de datos de la fuente primaria.
- Realizar la limpieza de los datos mediante las siguientes acciones en este orden:
    - Eliminación de columnas referentes a las variables que no son de nuestro interés.
    - Tratamiento de datos nulos, mediante el remplazo de estos por la media de la variable correspondiente.
    - Tratamiento de datos atípicos (outliers) mediante la detección de ellos con un boxplot y la eliminación mediante el método de cuartiles para definir límites.
- Creación de 4 datasets de acuerdo con los datos contenido en la variable 'room_type'.
- Obtener coeficientes de correlación y coeficientes de determinación de cada dataset.
- Escoger la mejor variable o mejores variables (correlación más fuerte) para realizar modelo de regresión lineal.
- Realizar modelo de regresión lineal.

Resultados.

A continuación, los resultados obtenidos en las 3 ciudades respecto a los coeficientes de correlación y el modelo final escogido que describe mejor el comportamiento de la variable 'número de reseñas'.

Ciudad 1: CDMX, México

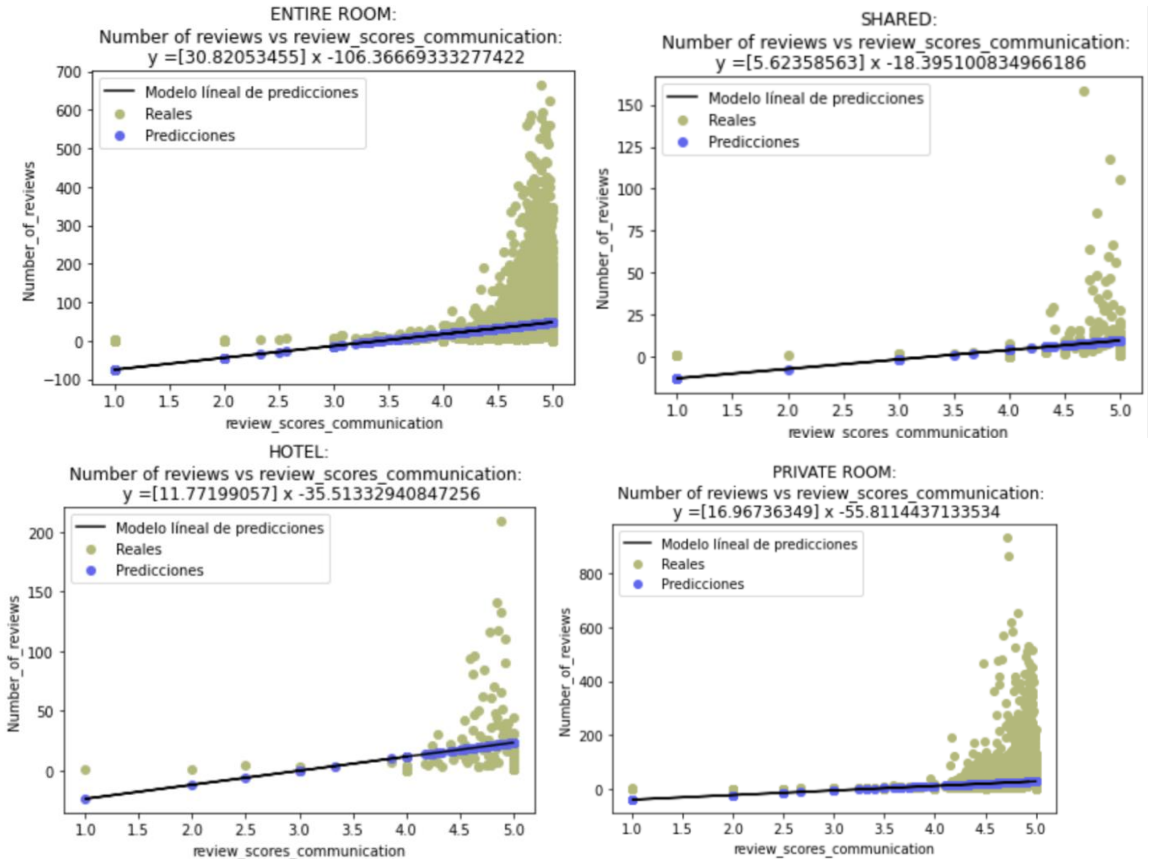
A. *Tabla de coeficientes de correlación y determinación: ''number_of_reviews'.*

   El primer número corresponde al coeficiente de determinación, y el segundo de correlación.

Variable/Room type	ENTIRE HOME/ APT	PRIVATE ROOM	HOTEL	SHARED ROOM
availability_365	0.0016 0.0404	0.0013 0.0366	0.0056 0.0749	0.0022 0.0467
review_scores_rating	0.0342 0.1851	0.034 0.1845	0.0494 0.2224	0.0419 0.2047
price	0.0003 0.0177	0.0004 0.0197	0.005 0.0707	0.0056 0.0745
host_acceptance_rate	0.0211 0.1456	0.0392 0.1981	0.0325 0.1803	0.0535 0.2313
review_scores_cleanliness	0.0405 0.2013	0.0427 0.2066	0.0412 0.2029	0.0501 0.2238
review_scores_communication	**0.0465** **0.2158**	**0.0438** **0.2094**	**0.0516** **0.2273**	**0.0573** **0.2394**

Observamos en color verde los coeficientes de correlación más altos de cada cuarto, por lo tanto, fue la variable escogida para realizar el modelo de regresión lineal.

B. Mejor modelo por tipo de cuarto.

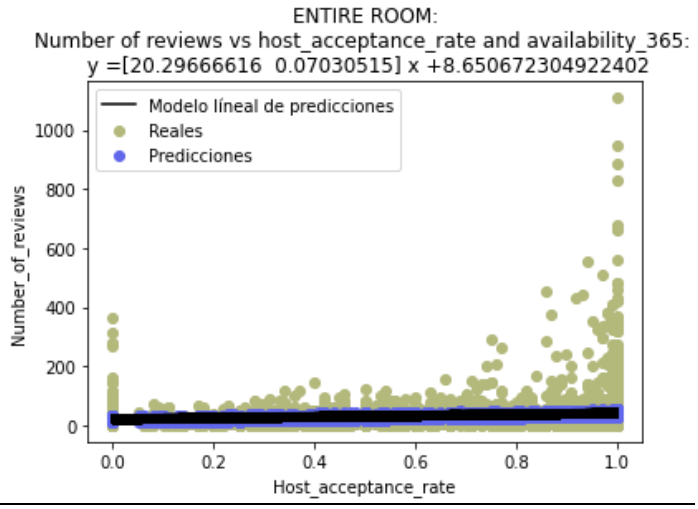


Ciudad 2: Berlín, Alemania

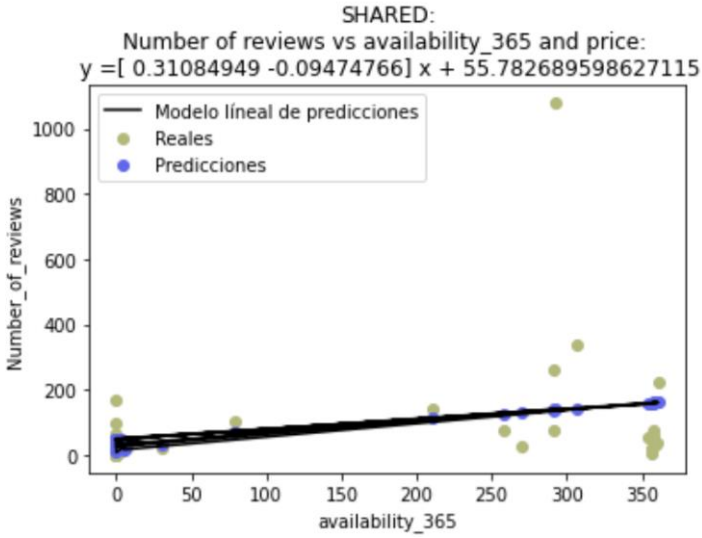
A. Tabla de correlación.

Variable/Room type	ENTIRE HOME/ APT	PRIVATE ROOM	HOTEL	SHARED ROOM
availability_365	**0.0205** **0.1432**	0.0027 0.0521	0.0233 0.1525	**0.0859** **0.2932**
review_scores_rating	0.0062 0.079	0.0235 0.1534	**0.0333** **0.1824**	0.0103 0.1013
price	0.0016 0.0395	**0.0393** **0.1982**	0.0072 0.0849	**0.0229** **0.1514**
host_acceptance_rate	**0.0227** **0.1505**	0.0277 0.1663	0.0125 0.112	0.0163 0.1275
review_scores_cleanliness	0.0114 0.1065	**0.0327** **0.1807**	0.0201 0.1419	0.0002 0.0148
review_scores_communication	0.0123 0.1111	0.0408 0.202	**0.0336** **0.1832**	0.0015 0.0387

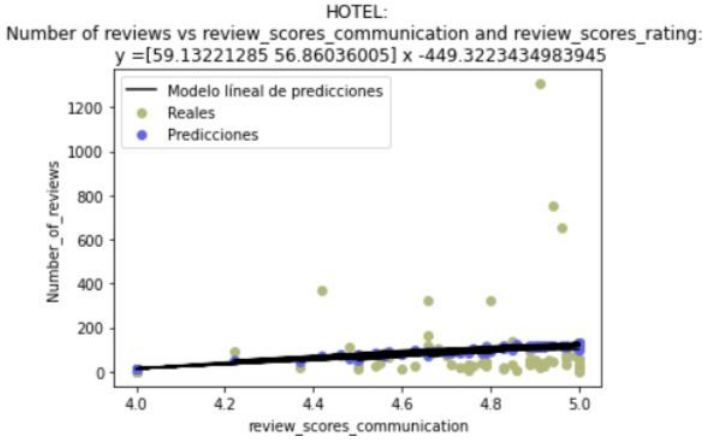
B. Mejor modelo.



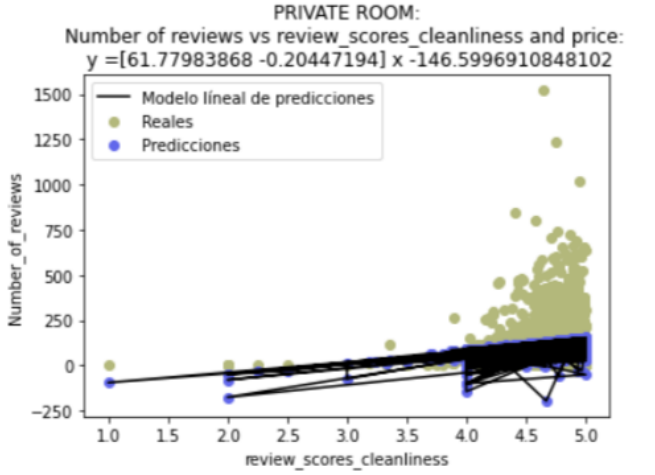
coeficiente de Determinación 0.03867754424911507
coeficiente de Correlación 0.1966660729488314



coeficiente de Determinación 0.08871481173641116
coeficiente de Correlación 0.2978503176704889



coeficiente de Determinación 0.035214315700188736
coeficiente de Correlación 0.18765477798390515



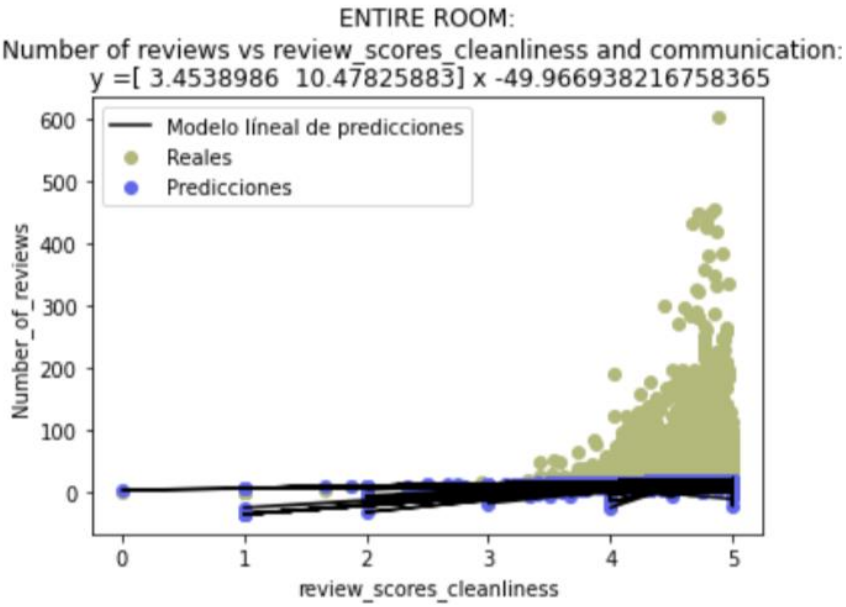
coeficiente de Determinación 0.06441630527685138
coeficiente de Correlación 0.2538036746716867

Ciudad 3: Hovedstaden, Dinamarca.

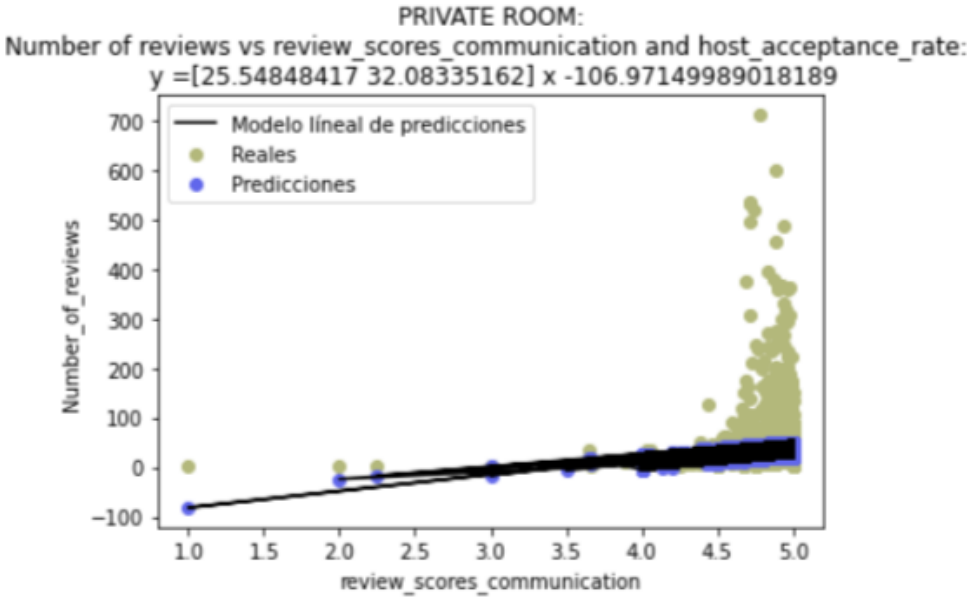
A. Tabla de correlación.

Variable/Room type	ENTIRE HOME/ APT	PRIVATE ROOM	HOTEL	SHARED ROOM
availability_365	0.0031 0.0553	0.0005 0.0217	0.005 0.0706	**0.1788** **0.4228**
review_scores_rating	0.0134 0.1159	0.014 0.1183	0.0018 0.0422	0.0333 0.1825
price	0.0001 0.0112	0.0007 0.0268	**0.2122** **0.4607**	0.0502 0.2241
host_acceptance_rate	0.0164 0.1281	**0.0423** **0.2057**	0.0019 0.0431	**0.2322** **0.4818**
review_scores_cleanliness	**0.0174** **0.1319**	0.0262 0.1619	**0.1343** **0.3664**	0.0189 0.1376
review_scores_communication	**0.0265** **0.1628**	**0.0306** **0.1748**	0.0044 0.0662	0.0497 0.2229

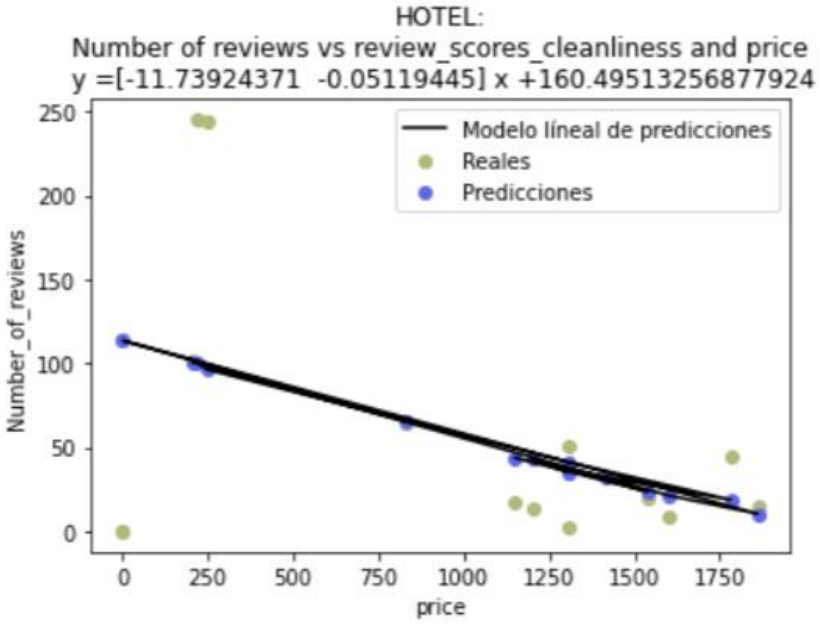
B. Mejor modelo.



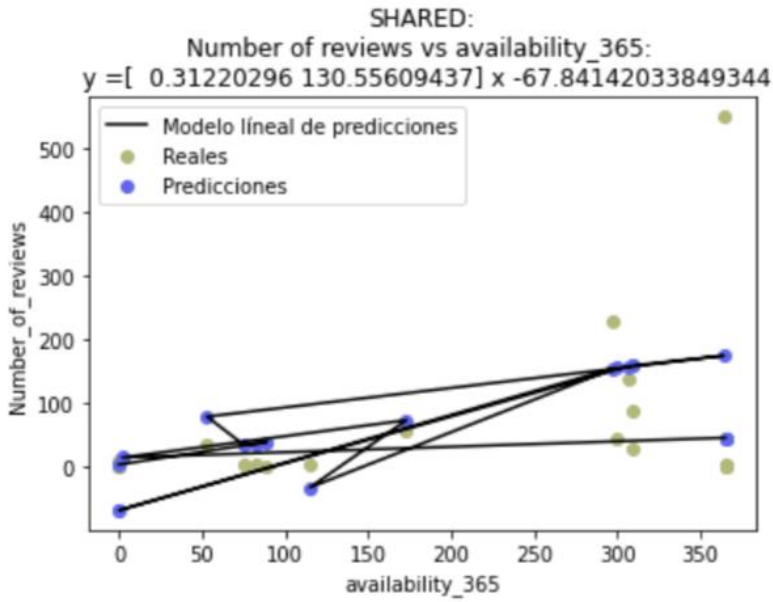
coeficiente de Determinación 0.028109376152052223
coeficiente de Correlación 0.1676585105267616



coeficiente de Determinación 0.06325307907928068
coeficiente de Correlación 0.25150164826354654



coeficiente de Determinación 0.21333200567299304
coeficiente de Correlación 0.4618787781149866




```
coeficiente de Determinación 0.3502830235260599
coeficiente de Correlación 0.5918471285104456
```

Interpretación de Resultados

Por Ciudad:

México:

Coeficientes

Observamos en la tabla de coeficientes para la Ciudad de México, que en general tenemos correlaciones muy débiles o casi nulas. Esto nos dice a priori que nuestra variable (determinada como dependiente) de número de reseñas no tiene una relación con las variables independientes que fijamos. Al mismo tiempo, nuestros coeficientes de determinación, igualmente muy bajos, nos habla de que la proporción de la varianza explicada es casi inexistente en comparación con la proporción de la varianza total. Esto lo veremos reflejado en los modelos.

Modelo

Para cada tipo de cuarto (hotel, private, apt, y shared) se realizó un modelo de regresión lineal simple, es decir con una variable (la de mayor correlación). Por tal motivo, la línea del modelo en la gráfica sigue una línea recta tratando de describir el modelo.

Para este caso de la ciudad de México, las correlaciones más altas tuvieron la misma variable de 'review_scores_communication' , con alrededor del coeficiente de 0.2.

Por lo tanto, concluimos que a pesar de que una misma variable se llevó las correlaciones más fuertes, estas siguen siendo débiles, con modelos que no describen el comportamiento de los datos.

Berlín:

Coeficientes

Al igual que en el caso de CDMX, en Berlín presentamos coeficientes de correlación y determinación muy bajos, incluso más bajos que en México, siendo el más alto de alrededor de 0.18. Esto nos dice que no existe ninguna relación entre nuestras variables y además que la proporción de varianza explicada por el modelo es baja en comparación con la Total.
A diferencia de CDMX que la variable coincida que era la misma para todos los tipos de cuarto, aquí ten emos diferentes variables dependiendo si se trata de HOTEL, SHARED, APT o PRIVATE.

Modelo

Para esta Ciudad, en vez de realizarse un modelo de regresión lineal simple, se realizó múltiple, es decir en vez que seleccionar 1 variable independiente, se escogieron las 2 con coeficientes de correlación más alto. Esta decisión provocó que los modelos presentarán una mejora, viéndose reflejado en coeficientes más altos que si hubieran estado con una sola variable. Sin embargo, siguen siendo modelos que no describen adecuadamente los datos.

Dinamarca:

Coeficientes

Para esta tercera Ciudad, ya tenemos coeficientes más altos, siendo el mayor de alrededor de 0.4, lo cual podemos considerar una correlación mediana. Además de que al igual que en Berlín también tenemos diferentes variables más altas dependiendo del tipo de cuarto del que se trate.

Modelo

En cuanto a los modelos, también vemos una mejora significativa al realizar modelos de regresión lineal múltiple con 2 variables, puesto que la proporción de la varianza de los datos explicada aumento, esto al igual que la relación que existe entre las variables. Esta relación la podemos atribuir a una causalidad existente o simplemente coincidencias en las correlaciones de los atributos.

Entre Ciudades.

Comparando los modelos entre Ciudades observamos que:

- Cada Ciudad varía los datos, viéndose reflejado en la selección de las variables es decir, la misma variable en CDMX, para todos los modelos, mientras que diferentes variables en las ciudades de Berlín y Dinamarca, donde en este último obtuvimos mejores resultados.
- Las correlaciones entre las variables seleccionadas no presentan en general relaciones, concluyendo que las variable ‘dependiente’ de número de reseñas no depende de las demás.
- En general las proporciones de varianza explicadas por las 3 es baja.

Conclusiones.

Pera concluir podemos destacar 2 puntos:
La importancia de realizar una limpieza y preparar bien los datos antes de realizar las operaciones o los modelos, ya que en gran medida del trabajo dependerá como tratamos a nuestros datos.

Ningún modelo es bueno por su correlación débil, por lo tanto, las gráficas de regresión lineal y múltiple no realizan correctamente predicciones y no describen bien los datos.
Sin embargo, el modelo de múltiple es mejor que el simple.

De las 3 ciudades tiene mejores resultados de correlaciones y modelos la de en Dinamarca. Aquí si puede existir una correlación real, mediana. Estos nos dicen que, en cada ciudad, van a ser diferentes los datos y por lo tanto tendremos resultados diferentes.