María Luisa Benítez Guerrero
A01706447

Reporte comparativo – Actividad 1

Con el fin de analizar el comportamiento de tres ciudades con modelos de regresión lineal simple y múltiple, se eligieron los datos de la Ciudad de México para el modelo de regresión lineal simple, y los datos de Melbourne y Milán para el modelo de regresión lineal múltiple.

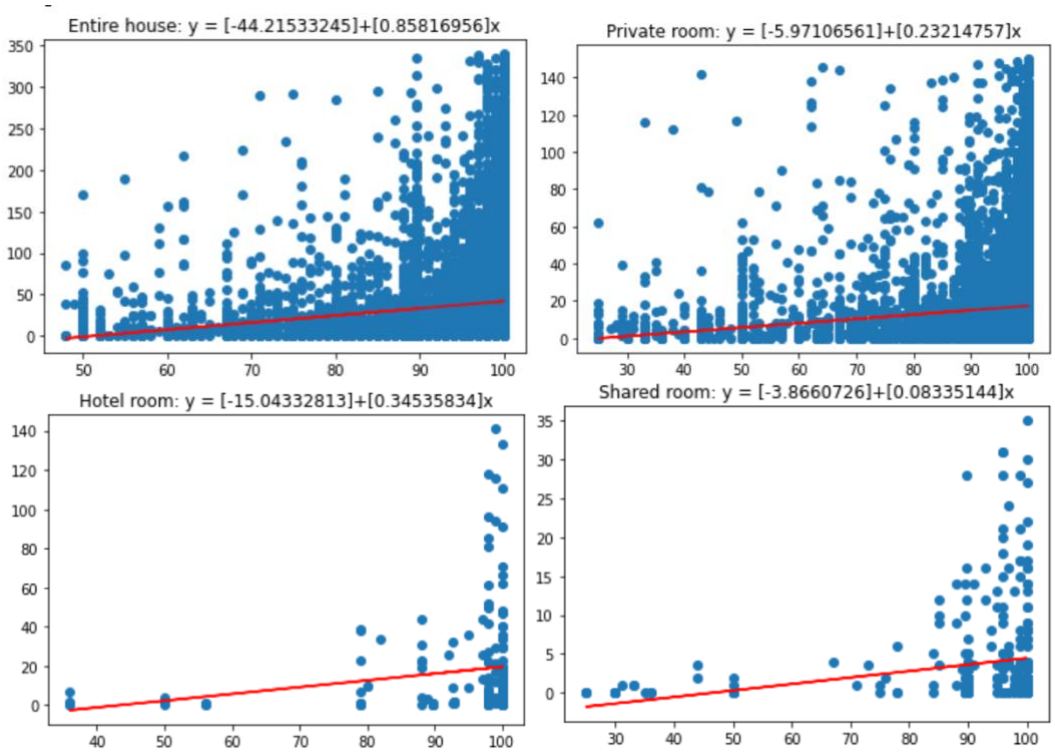
En los tres casos se inició preparando y limpiando los datos, quitando caracteres que complicarían el análisis, así como realizando las acciones de preprocesamiento necesarias con datos nulos y con los outliers. Posteriormente se dividieron todas las bases de datos en cuatro dataframes diferentes, cada uno por el tipo de cuarto que existía en las bases de datos.

Para el caso de la Ciudad de México, se analizó la correlación que existía en cada tipo de habitación respecto a las variables 'price','host_acceptance_rate','availability_365','review_scores_rating','review_scores_cleanliness' y 'review_scores_communication' como variables predictoras y 'number_of_reviews' como variable resultado, para después realizar un modelo de regresión lineal y evaluar su correlación con cada uno, obteniendo los siguientes resultados:

```
Entire home/apt
price 0.004444278399522661
host_acceptance_rate 0.019563187378069835
availability_365 0.003913286990574361
review_scores_rating 0.0038703930319675672
review_scores_cleanliness 0.004296031765542274
review_scores_communication 0.005560387637971065

Private room
price 0.005389172949229026
host_acceptance_rate 0.016215860030568074
availability_365 0.008690376476888528
review_scores_rating 0.00787036131415253
review_scores_cleanliness 0.0026858856315504775
review_scores_communication 0.001932456755003531

Shared room
price 0.011165025313762289
host_acceptance_rate 0.044839979816536424
availability_365 0.02569869846419659
review_scores_rating 0.00016934429524950723
review_scores_cleanliness 0.006785791502896599
review_scores_communication 0.0015192615835777357

Hotel room
price 0.012826996900350163
host_acceptance_rate 0.03303149877455891
availability_365 5.675749483446957e-05
review_scores_rating 0.009656069340207574
review_scores_cleanliness 0.00013081591699626305
review_scores_communication 0.0015616242567080274
```

Y después se creó el modelo matemático que describiera de mejor manera el número de reseñas para cada tipo de alojamiento, utilizando la variable con mayor correlación, la cual fue ‘host_acceptance_rate' en todos los casos.

Cabe mencionar que estas correlaciones eran extremadamente bajas, por lo que desde aquí no se espera tener muy buenos resultados.

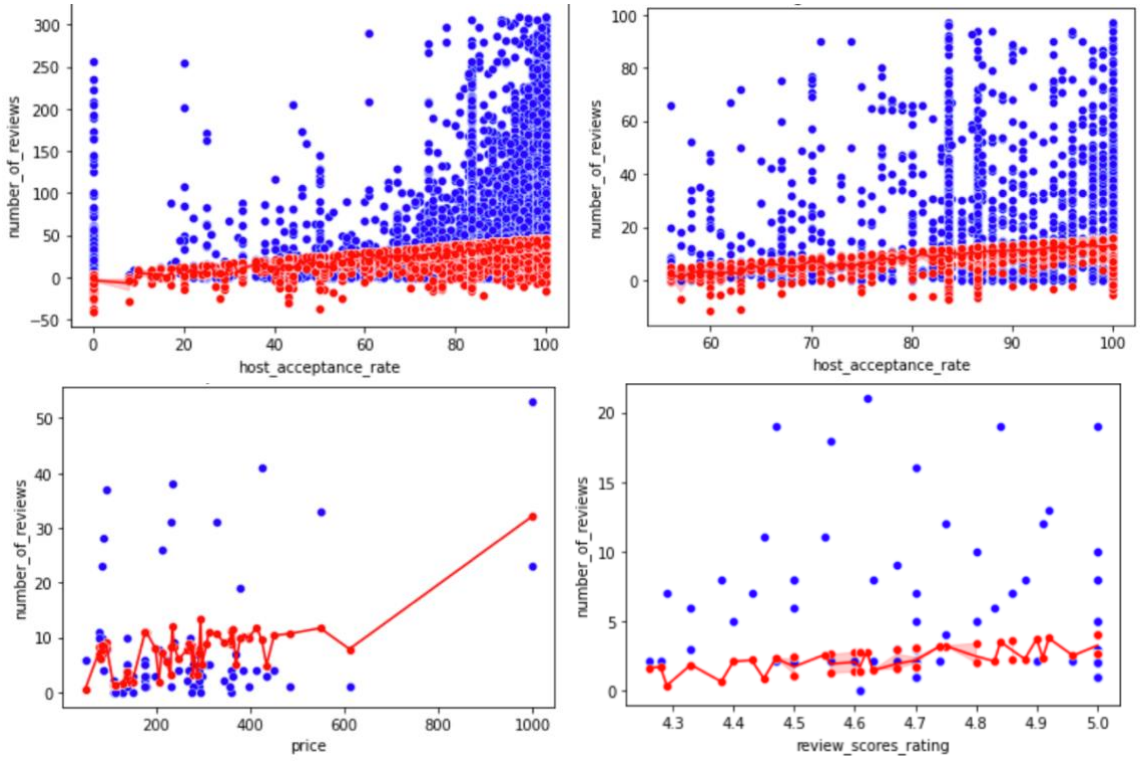


Para el caso de las ciudades de Melbourne y Milán, primero se obtuvo la correlación en cada base de datos para conocer qué variables tenían la mayor correlación con el número de reseñas, y con base en eso se elegían las variables que se evaluarían en el modelo de regresión múltiple, obteniendo los modelos siguientes:

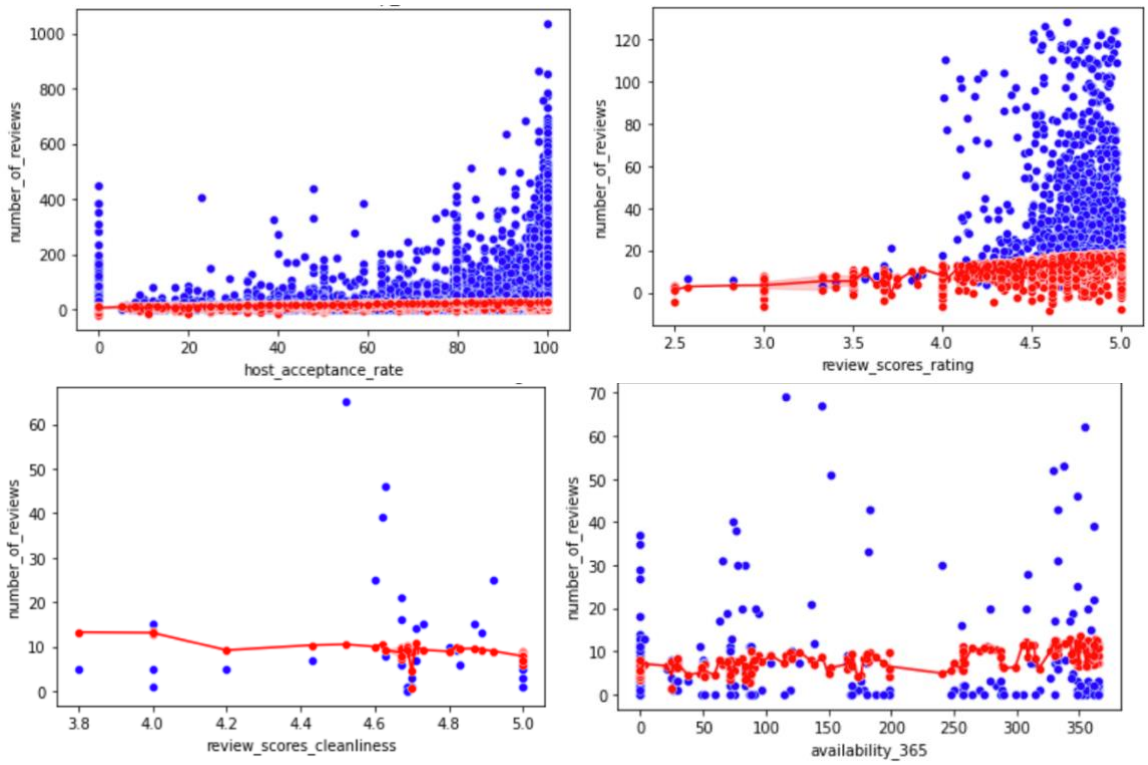
Ciudad	Tipo de cuarto	Modelos	Correlación
Melbourne	Entire home	y = -127.599 + 0.4311(accep_rate) + 18.1339(rating) + 8.1096(cleanliness)	0.0482
Melbourne	Private room	y = -52.4698 + 7.3446(rating) + 1.3795(cleanliness) + 0.2456(accep_rate)	0.0417
Melbourne	Hotel room	y = 11.9845(price) + 0.01951(available_365) -0.1443 (accep_rate)	0.2291
Melbourne	Shared room	y = 6969.4609 + 3.2465 (rating) - 83.5126 (accep_rate)	0.0526
Milán	Entire home	y = 9.0769 + 0.2094 (accep_rate) - 0.03608 (price)	0.0400

Milán	Private room	y = -16.1034 + 0.08426 (accept) - 0.0479 (price) + 5.6471 (rating)	0.0389
Milán	Hotel room	y = 30.0234 - 7.5501(cleanliness) + 3.4022(rating) - 0.00505 (availabil)	0.0277
Milán	Shared room	y = 10.3190 + 0.01601(availability_365) + 0.05111 (accep_rate)	0.0526

Las gráficas se presentan en el orden de la tabla, empezando por Melbourne:



Y después Milán:



Finalmente, se obtuvieron los coeficientes de determinación y correlación de cada uno de los modelos, quedando así:

Melbourne

	Room type	Coef_deter	Coef_correl
0	Entire room	0.019563	0.139868
1	Private room	0.016216	0.127342
2	Hotel room	0.033031	0.181746
3	Shared room	0.044840	0.211755

Milan

	Room type	Coef_deter	Coef_correl
0	Entire room	0.039979	0.199947
1	Private room	0.038906	0.197246
2	Hotel room	0.027652	0.166290
3	Shared room	0.039447	0.198614

	Room type	Coef_deter	Coef_correl
0	Entire room	0.048207	0.219561
1	Private room	0.041650	0.204084
2	Hotel room	0.229065	0.478607
3	Shared room	0.052580	0.229303

Con estos resultados podemos observar que no existe diferencia significativa entre el uso de regresión lineal y regresión múltiple, esto debido a que desde el principio la correlación entre las variables de predicción y el

número de reseñas no es nada fuerte. Se les hicieron modificaciones a los datos para prepararlos y tenerlos lo más completos posibles, pero eso no fue suficiente para crear modelos confiables. El modelo con un coeficiente de correlación más alto es el de hotel room de Melbourne, con 47%, pero aun así no es suficiente para alcanzar un nivel confiable de correlación.

En cuanto a la diferencia del modelo de regresión lineal simple y regresión lineal múltiple, tampoco podemos observar gran diferencia entre sus coeficientes, los cuales todos están en un rango de 0.12 a 0.23 (exceptuando el 0.47 mencionado anteriormente), demostrando que cuando se cuenta con datos que no tienen correlación entre sí, no se puede hacer mucho por ellos para contar con modelos confiables.