**Actividad 6.1: Regresión lineal simple**

Miguel Saúl Fernández Avalos  
A01707491

Se Realizaron los procedimientos de eliminación de NA y outliers iguales que para la actividad anterior por lo que no se hará énfasis en esta parte. Para el análisis de los modelos, se eligió analizar para las tres ciudades el tipo de habitación Entire home/apt y Private room, ya que para las tres ciudades éstos eran los tipos de cuarto más comunes, por lo que se tenía más información para generar los modelos.

Después de tener los filtros con la información específica, se empezó con la generación de modelos, teniendo las siguientes variables independientes y dependientes:

1. Independiente: host\_response\_rate

Dependiente: host\_acceptance\_rate

1. Independiente: review\_scores\_cleanliness

Dependiente: review\_scores\_location

1. Independiente: price

Dependiente: host\_acceptance\_rate

1. Independiente: number\_of\_reviews

Dependiente: availability\_365

1. Independiente: number\_of\_reviews

Dependiente: host\_acceptance\_rate

1. Independiente: review\_scores\_communication

Dependiente: reviews\_per\_month

Al ser 2 tipos de habitaciones, se tendrían que generar 12 modelos, para facilitar su generación se realizó un ciclo donde se implementaría el modelo de regresión linear, con la variable independiente y dependiente siendo cambiados por las variables mencionadas, para posteriormente imprimir la ecuación del modelo, con el método \_dict\_ se extrajo el coeficiente del modelo y la intersección obtenida del modelo para generar la ecuación.

Adicionalmente, se generó el score del modelo, que explica su coeficiente de determinación, así como el coeficiente de correlación que es precisamente la raíz cuadrada del coeficiente de determinación, ambos valores se guardaron en listas para posteriormente integrarlos a un dataframe para facilitar la comparación de resultados.

Los cuales son los siguientes, del 0 al 5 corresponden al tipo de cuarto Entire home/apt y del 6 en adelante corresponden a Private room para las tres ciudades.

D.F. Girona

|  | **Determinacion** | **Correlacion** |
| --- | --- | --- |
| **0** | 0.120748 | 0.347488 |
| **1** | 0.043622 | 0.208858 |
| **2** | 0.000008 | 0.002841 |
| **3** | 0.004755 | 0.068959 |
| **4** | 0.021059 | 0.145116 |
| **5** | 0.013526 | 0.116303 |
| **6** | 0.119415 | 0.345565 |
| **7** | 0.073283 | 0.270709 |
| **8** | 0.000018 | 0.004287 |
| **9** | 0.004365 | 0.066068 |
| **10** | 0.021823 | 0.147726 |
| **11** | 0.039974 | 0.199935 |

|  | **Determinacion** | **Correlacion** |
| --- | --- | --- |
| **0** | 0.079245 | 0.281506 |
| **1** | 0.028012 | 0.167368 |
| **2** | 0.000107 | 0.010342 |
| **3** | 0.000146 | 0.012090 |
| **4** | 0.000005 | 0.002227 |
| **5** | 0.005453 | 0.073845 |
| **6** | 0.127289 | 0.356776 |
| **7** | 0.030964 | 0.175967 |
| **8** | 0.002145 | 0.046317 |
| **9** | 0.000059 | 0.007652 |
| **10** | 0.007257 | 0.085191 |
| **11** | 0.036380 | 0.190735 |

California

|  | **Determinacion** | **Correlacion** |
| --- | --- | --- |
| **0** | 0.044125 | 0.210060 |
| **1** | 0.037154 | 0.192753 |
| **2** | 0.006982 | 0.083560 |
| **3** | 0.008143 | 0.090240 |
| **4** | 0.009984 | 0.099920 |
| **5** | 0.035047 | 0.187209 |
| **6** | 0.018779 | 0.137038 |
| **7** | 0.058992 | 0.242883 |
| **8** | 0.002818 | 0.053081 |
| **9** | 0.003585 | 0.059877 |
| **10** | 0.020265 | 0.142354 |
| **11** | 0.031611 | 0.177796 |

Para el DF, en general, no se observa en ningún caso correlación negativa, aunque tampoco correlaciones medias o superiores. Donde se encuentra la correlación es en el modelo de host\_acceptance\_rate vs host\_response\_rate para ambos tipos de cuartos, con poco más de 0.34 de correlación. Por otro lado, se pude observar que para host\_acceptance\_rate vs Price hay una correlación casi nula.

Para Girona observamos nuevamente que la mayor correlación se encuentra en el modelo generado para host\_acceptance\_rate vs host\_response\_rate, aunque en este caso donde se observa correlación casi nula es en el modelo correspondiente a host\_acceptance\_rate vs number\_of\_reviews Entire home/apt y availability\_365 vs number\_of\_reviews para Private room.

Finalmente, para California, el modelo con mayor correlación para Entire home/apt corresponde a host\_acceptance\_rate vs host\_response\_rate, mientras que para Private room es el modelo de review\_scores\_location vs review\_scores\_cleanliness, en general se puede observar que a pesar de tener la mejor correlación, ésta es menor que los modelos de Girona y DF por más de 0.1. Por otro lado, los modelos con menor correlación tienen el efecto contrario que con DF y Girona, son superiores a los de las otras dos ciudades, aunque tampoco llegan a ser considerados siquiera como correlación débil.

Considerando que ni siquiera hay una correlación fuerte entre las variables, no vale la pena analizar si existe causalidad, se puede observar que no sería muy explicativo tratar de usar estos modelos para predicciones.