

## Evaluación: Actividad 6.1 (Regresión Lineal Simple)

En esta primera etapa, primeramente empecé cargando todos mis archivos de csv. Para que mi modelo de regresión funcionara correctamente tuve que limpiarlo y rellenar datos nulos. Para mis datos string, los rellené con un before fill y forward fill y para mis datos numéricos rellene con una mediana. Posteriormente, apliqué reducción de valores extremos por medio del rango intercuartil, para reducir el ruido y valores demasiado dispersos. Este paso los repetí para los 3 data frames utilizados.

Posteriormente, pasé a la parte de correlación. Para cada uno de mis dataframes ya preprocesados, pasé datos string a datos que deberían ser enteros para poder trabajar con ellos, para después seleccionar solo las filas que tuvieran como característica 'Entire home/apt' y 'Private room'.

También, aislamos las variables de cada tipo de cuarto como variables dependientes o independientes según sea el caso.

A continuación, dibujé solo con las dos primeras variables, en una gráfica de dispersión, para poder ver la correlación que tenían.

Ya teniendo nuestros datos listos, se llevaron a cabo las regresiones en las cuales tuvimos los siguientes resultados:

- México
  - host\_acceptance\_rate vs host\_response\_rate: En esta estamos midiendo el efecto que tiene que un host responda rápido en la calificación final que recibe. Podemos notar que en ambos, tienen una baja correlación, podríamos decir que el host que responde rápido no necesariamente recibirá una buena calificación. También, nuestro modelo nos da un R2 muy bajo en ambos cuartos, dándonos a entender que la fuerza de la relación lineal es baja. No podríamos considerar esta variable únicamente como explicativa para la calificación del host en México.
    - Entire home/apt:
      - Correlación: 0.35
      - Determinación: 0.12
    - Private room:
      - Correlación: 0.35
      - Determinación: 0.12
  - review\_scores\_location vs review\_scores\_cleanliness: En esta prueba medimos cuál es el efecto de la calificación en limpieza en la calificación por localización. Hasta al mencionar lo que mide, la lógica nos dice que no tendrían mucho que ver, e igualmente nuestro modelo nos indica que no hay correlación significativa ni la variación de la variable independiente explica lo suficiente para asegurar con cierto porcentaje de confianza variable dependiente.
    - Entire home/apt:
      - Correlación: 0.20

- Determinación: 0.04
  - Private room:
    - Correlación: 0.27
    - Determinación: 0.07
  - host\_acceptance\_rate vs price: Aquí medimos cómo afecta el precio del alojamiento en la tasa de aceptación del anfitrión. De igual manera en ambos cuartos tenemos coeficientes bajos, dándonos a entender que no podríamos considerar esta variable únicamente como explicativa para la tasa de aceptación del host en México.
    - Entire home/apt:
      - Correlación: 0.002
      - Determinación: 0.000008
    - Private room:
      - Correlación: 0.004
      - Determinación: 0.000018
  - availability\_365 vs number\_of\_reviews: Aquí medimos cómo el número de reseñas afecta la ocupación del alojamiento. De igual manera encontramos que no tienen nada que ver con la ocupación del alojamiento.
    - Entire home/apt:
      - Correlación: 0.06
      - Determinación: 0.004
    - Private room:
      - Correlación: 0.06
      - Determinación: 0.004
  - host\_acceptance\_rate vs number\_of\_reviews: Aquí medimos el posible efecto que tiene el número de reseñas en la tasa de aceptación del host. Aquí las reseñas, no tienen nada que ver con la aceptación que tiene el anfitrión en ninguno de los dos tipos de cuarto en México.
    - Entire home/apt:
      - Correlación: 0.14
      - Determinación: 0.02
    - Private room:
      - Correlación: 0.14
      - Determinación: 0.02
  - reviews\_per\_month vs review\_scores\_communication: Se mide el efecto de la calificación en comunicación en el número de reseñas por mes. Nuestro modelo encontró que no hay relación entre estas dos variables
    - Entire home/apt:
      - Correlación: 0.002
      - Determinación: 0.000008
    - Private room:
      - Correlación: 0.11
      - Determinación: 0.01
- California: En general, California tiene datos un poco más altos pero por décimas. Aún así, no encontramos ninguna correlación entre las variables ni que únicamente esa, fuera explicativa para la dependiente. Se repite el mismo veredicto que en México para cada una de las regresiones.
  - host\_acceptance\_rate vs host\_response\_rate:

- Entire home/apt:
  - Correlación: 0.21
  - Determinación: 0.04
- Private room:
  - Correlación: 0.13
  - Determinación: 0.02
- review\_scores\_location vs review\_scores\_cleanliness:
  - Entire home/apt:
    - Correlación: 0.29
    - Determinación: 0.04
  - Private room:
    - Correlación: 0.24
    - Determinación: 0.06
- host\_acceptance\_rate vs price:
  - Entire home/apt:
    - Correlación: 0.08
    - Determinación: 0.006
  - Private room:
    - Correlación: 0.05
    - Determinación: 0.002
- availability\_365 vs number\_of\_reviews: o.
  - Entire home/apt:
    - Correlación: 0.09
    - Determinación: 0.008
  - Private room:
    - Correlación: 0.06
    - Determinación: 0.003
- host\_acceptance\_rate vs number\_of\_reviews:
  - Entire home/apt:
    - Correlación: 0.09
    - Determinación: 0.008
  - Private room:
    - Correlación: 0.14
    - Determinación: 0.02
- reviews\_per\_month vs review\_scores\_communication:
  - Entire home/apt:
    - Correlación: 0.18
    - Determinación: 0.03
  - Private room:
    - Correlación: 0.18
    - Determinación: 0.03
- Girona: En general, en Girona, tampoco vemos mucha relación entre nuestras variables. Son muy parecidas al caso de México.
  - host\_acceptance\_rate vs host\_response\_rate:
    - Entire home/apt:
      - Correlación: 0.28
      - Determinación: 0.08
    - Private room:

- Correlación: 0.36
- Determinación: 0.12
- review\_scores\_location vs review\_scores\_cleanliness:
  - Entire home/apt:
    - Correlación: 0.16
    - Determinación: 0.03
  - Private room:
    - Correlación: 0.17
    - Determinación: 0.03
- host\_acceptance\_rate vs price:
  - Entire home/apt:
    - Correlación: 0.01
    - Determinación: 0.0001
  - Private room:
    - Correlación: 0.04
    - Determinación: 0.002
- availability\_365 vs number\_of\_reviews: o.
  - Entire home/apt:
    - Correlación: 0.01
    - Determinación: 0.0001
  - Private room:
    - Correlación: 0.007
    - Determinación: 0.00006
- host\_acceptance\_rate vs number\_of\_reviews:
  - Entire home/apt:
    - Correlación: 0.002
    - Determinación: 0.000005
  - Private room:
    - Correlación: 0.085
    - Determinación: 0.007
- reviews\_per\_month vs review\_scores\_communication:
  - Entire home/apt:
    - Correlación: 0.07
    - Determinación: 0.005
  - Private room:
    - Correlación: 0.19
    - Determinación: 0.03

Finalmente podemos concluir que los coeficientes de determinación son demasiados bajos para considerar el modelo matemático como explicativo para cada uno de los casos vistos, en todas las ciudades. Tal vez se debería buscar otra variable explicativa para cada uno de los casos, o pasar a modelos más complejos. Puesto que suena un poco más lógico que los factores de una calificación está basada en varios aspectos más que en solo 1.