INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

ACTIVIDAD 6-1

Nombre del alumno:

José María Colombres Elguea | A01734153

Análisis de datos y herramientas de inteligencia artificial I

(Grupo 101)

Profesores: Candy Yuridiana Alemán Muñoz, Alfredo García Suárez

Fabiola Díaz Nieto & Francisco Javier Navarro Barrón

Fecha de entrega: 04 de mayo de 2023

ACTIVIDAD 6-1

Para realizar el análisis de regresión se utilizaron dos tipos de filtro para la columna “room_type” que fueron aplicados a las variables. En el filtro 1 se separó por “Entire home/apt", mientras que en el filtro 2 se separó por “Private room”. A continuación se muestra una tabla de todos los coeficientes de determinación y correlación obtenidos para cada tipo de habitación elegida.

DF | México

DF_México | Dataframe			
Columnas	**Filtro**	**% Determinación**	**% Correlación**
host_acceptance_rate vs host_response_rate	1	0.120	0.347
	2	0.118	0.344
host_acceptance_rate vs price	1	2.275	0.004
	2	2.650	0.005
host_acceptance_rate vs number_of_reviews	1	0.011	0.106
	2	0.038	0.197
review_scores_location vs review_scores_cleanliness	1	0.048	0.220
	2	0.075	0.274
availability_365 vs number_of_reviews	1	0.001	0.036
	2	0.0007	0.027
reviews_per_month vs review_scores_communication	1	0.016	0.129
	2	0.040	0.200

Al analizar los resultados obtenidos para el dataframe de México DF, se observa que ningún coeficiente de correlación sobrepasa los 35 puntos porcentuales, lo que indica que las variables analizadas tienen una correlación positiva débil. El conteo es el siguiente:

- **Correlación positiva débil (+0.10):** 8 | **Correlación nula: (0):** 4

El porcentaje de correlación más alto para ambos filtros se obtuvo al comparar host_acceptance_rate vs host_response_rate. Esto fue anticipable debido a que en un análisis

cualitativo, estas columnas están directamente relacionadas, ya que si un host tiene un porcentaje alto de respuesta, también lo tendrá en aceptación. De igual forma el porcentaje de correlación más pequeño se dió al utilizar las columnas de host_acceptance_rate vs price, debido a que manejan diferentes tipos de datos, y además son de categorías distintas.

California | USA

California_USA | Dataframe			
Columnas	**Filtro**	**% Determinación**	**% Correlación**
host_acceptance_rate vs host_response_rate	1	0.080	0.283
	2	0.085	0.292
host_acceptance_rate vs price	1	0.001	0.034
	2	0.002	0.049
host_acceptance_rate vs number_of_reviews	1	0.012	0.109
	2	0.026	0.163
review_scores_location vs review_scores_cleanliness	1	0.049	0.223
	2	0.052	0.228
availability_365 vs number_of_reviews	1	0.010	0.104
	2	0.011	0.107
reviews_per_month vs review_scores_communication	1	0.043	0.208
	2	0.038	0.195

Al analizar los resultados obtenidos para el dataframe de California, se observa que ningún coeficiente de correlación sobrepasa los 30 puntos porcentuales, lo que indica que las variables analizadas tienen una correlación positiva débil. El conteo es el siguiente:

- **Correlación positiva débil (+0.10):** 10 | **Correlación nula: (0):** 2

El porcentaje de correlación más alto para ambos filtros se obtuvo al comparar host_acceptance_rate vs host_response_rate. Esto fue anticipable debido a que en un análisis cualitativo, estas columnas están directamente relacionadas, ya que si un host tiene un porcentaje alto de respuesta, también lo tendrá en aceptación. De igual forma el porcentaje de correlación más pequeño se dió al utilizar las columnas de host_acceptance_rate vs price, debido a que manejan diferentes tipos de datos, y además son de categorías distintas.

Comparativa

Columnas	F	DF_México | Dataframe		California_USA | Dataframe	
		% Determinación	% Correlación	% Determinación	% Correlación
host_acceptance_rate vs host_response_rate	1	0.120	0.347	0.080	0.283
	2	0.118	0.344	0.085	0.292
host_acceptance_rate vs price	1	2.275	0.004	0.001	0.034
	2	2.650	0.005	0.002	0.049
host_acceptance_rate vs number_of_reviews	1	0.011	0.106	0.012	0.109
	2	0.038	0.197	0.026	0.163
review_scores_location vs review_scores_cleanliness	1	0.048	0.220	0.049	0.223
	2	0.075	0.274	0.052	0.228
availability_365 vs number_of_reviews	1	0.001	0.036	0.010	0.104
	2	0.0007	0.027	0.011	0.107
reviews_per_month vs review_scores_communication	1	0.016	0.129	0.043	0.208
	2	0.040	0.200	0.038	0.195

Se puede observar que se obtuvieron mejores porcentajes de correlación con el dataframe de California. Esto puede explicarse debido a que tiene mucho menos registros que el de México, y además no presentó tantos registros nulos en la primera parte de limpieza. Por otro lado al analizar los promedios de los filtros por df se encontró lo siguiente:

DF | México

- Promedio correlación filtro 1: 0.1140
- Promedio correlación filtro 2: 0.1745

California | USA

- Promedio correlación filtro 1: 0.1601
- Promedio correlación filtro 2: 0.1723

El filtro 2 (que busca los valores de Private room) tiene una mayor correlación en promedio que el filtro 1 en ambos archivos.