



Tecnológico de Monterrey

Gestión de proyectos de plataformas Tecnológicas

Actividad 2

Profesor: Alfredo García Suárez

Manlio Fabio Rivera Pérez II A01734797

Gabriela Cantú Sánchez II A01735256

Análisis exploratorio

El objetivo de esta actividad fue realizar un análisis exploratorio tomando en cuenta diferentes países para poder ser comparados con México. Asimismo, se utilizaron diferentes plataformas de Inteligencia de Negocios para poder comparar su desempeño en diferentes tareas. Se decidió trabajar con los países México, Singapur y Londres, ya que deseamos que los países fueran diferentes para obtener hallazgos significativos de culturas, idiomas y continentes diferentes. Antes que nada, se llevaron a cabo acciones de preprocesamiento como la limpieza de valores nulos y de valores atípicos.

En la primera parte del análisis se llevaron a cabo estas acciones utilizando Python. El mismo código fue aplicado para las 3 bases de datos, por ende, se utilizaron las mismas acciones para rellenar o cambiar valores en las bases de datos de los 3 países. A partir de lo anterior, fue posible cargar las bases de datos a la plataforma de Minitab y hacer comparaciones en cuanto a la correlación que existe en cada tipo de habitación respecto a ciertas variables. Es importante recalcar que para la limpieza de valores nulos y atípicos se tomaron en cuenta los criterios o las columnas utilizadas para que los valores fueran reemplazados por valores acordes a la columna a sustituir.

Las distintas bases de datos contienen información sobre el desempeño de la plataforma de Airbnb en cada uno de los países seleccionados. Dentro de estas bases de datos se encuentran los tipos de habitaciones rentadas, su precio, las noches mínimas y máximas y más. Lo anterior será de utilidad ya que permitirá conocer los factores críticos para cada país y detectar si hay un patrón que se pueda extender hacia México.

Lo primero que se realizó en la plataforma de Minitab fue filtrar los datos limpios por tipo de habitación, como resultado se obtuvo una hoja para el hogar completo, una para el hotel, una para una habitación privada y una para una habitación compartida. Al tener lo anterior se analizó la correlación que existe entre el precio y alguna otra variable para cada habitación. Las variables que se analizaron con precio fueron; noches mínimas, número de reseñas, número de ofertas del anfitrión, disponibilidad 365, reseñas por mes y número de reseñas ltm. Con base a lo anterior se obtuvieron los siguientes resultados:

México

Con el uso de Minitab se obtuvieron los coeficientes de determinación, es decir, la R^2 . Debido a que el coeficiente de correlación es R, se obtuvo la raíz cuadrada del coeficiente de determinación. Es importante mencionar que Minitab arroja el resultado como porcentaje, por ende, en los 3 países y para cada habitación se dividió entre 100. El coeficiente de correlación es útil para medir el grado de correlación que existe entre las variables. Este coeficiente puede ser positivo o negativo, indicando el tipo de relación entre las variables. Asimismo, puede ser débil, moderadamente fuerte o fuerte, indicando la fortaleza de la relación entre las variables. Entre más se acerque a 1 el coeficiente, más fuerte es la relación entre las variables.

Como se puede observar en la siguiente tabla, los coeficientes de correlación fueron positivos. Esto nos permite entender que a medida de que una de las variables incrementa o decrementa, el precio generalmente también incrementa o decrementa. De igual manera, se puede observar que los coeficientes son muy chicos. Esto indica que las variables no tienen una correlación muy fuerte, de hecho en todos los casos sería considerada débil. En el caso del hogar completo, se encontró una correlación más fuerte entre el precio y la variable “calculated_host_listings_count”. Para el hotel, la correlación más alta fue entre precio y “availability_365”, para cuarto privado la correlación más alta fue entre precio y “calculated_host_listings_count”. Por último, para cuarto compartido, la correlación más alta fue entre precio y “calculated_host_listings_count”.

Es interesante analizar que para la mayoría de los tipos de habitaciones la correlación más fuerte fue entre precio y “calculated_host_listings_count”. Esto podría indicar que entre más propiedades tiene un anfitrión en airbnb más elevado es el precio que ofertan. A pesar de no ser utilizado en todos los casos, el número de reseñas también tuvo una fuerte correlación con el precio en la mayoría de los tipos de propiedades. De cualquier manera, es importante recalcar que la correlación no indica causalidad. Utilizando como ejemplo el análisis de la relación positiva entre variables, el hecho de que el número mínimo de noches aumente no implica que el precio también aumentará. Es decir, el número mínimo de noches no es causa del precio.

Por otra parte, se generó una regresión múltiple utilizando las 3 variables con la correlación más fuerte utilizando el precio como la variable endógena. Al utilizar más variables, teóricamente sería posible describir de mejor manera el precio para cada tipo de habitación. Es importante observar que para cada tipo de habitación se utilizaron combinación diferentes de variables. Lo anterior nos permite hacer distinciones de factores críticos para cada tipo de habitación. Por ejemplo, la disponibilidad de 365 no tiene un coeficiente de correlación tan alto en los hogares completos y esto se puede deber a que son más costosas o menos prácticas.

Coeficientes de correlación							
MEXICO	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	Múltiple
Entire home	0.03	0.06403124237	0.01414213562	0.1004987562	0.02449489743	0.06	0.1268857754
Hotel	0.1090871211	0.1208304597	0.05099019514	0.08	0.146628783	0.1053565375	0.2161018278
Private room	0.09643650761	0.07874007874	0.0331662479	0.1352774926	0.08185352772	0.03605551275	0.1734935157
Shared room	0.03464101615	0.07549834435	0.1170469991	0.1459451952	0.06164414003	0.07348469228	0.2022374842

Coeficientes de determinación							
MEXICO	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	Múltiple
Entire home	0.0009	0.0041	0.0002	0.0101	0.0006	0.0036	0.0161
Hotel	0.0119	0.0146	0.0026	0.0064	0.0215	0.0111	0.0467
Private room	0.0093	0.0062	0.0011	0.0183	0.0067	0.0013	0.0301
Shared room	0.0012	0.0057	0.0137	0.0213	0.0038	0.0054	0.0409

Al obtener el coeficiente de determinación de la regresión múltiple para cada tipo de habitación fue posible visualizar que es el modelo matemático que mejor describe la variable de precio para todos los casos. Por ende, se elaboraron las siguientes ecuaciones de regresión ya que son el mejor modelo matemático para describir la variable de precio para cada caso;

Entire home

- $\text{price} = 1606.5 - 3.914 \text{ number_of_reviews_ltm} - 1.741 \text{ number_of_reviews} + 6.422 \text{ calculated_host_listings_count}$

Hotel room

- $\text{price} = 1467 - 284 \text{ minimum_nights} - 5.85 \text{ number_of_reviews} + 2.056 \text{ availability_365}$

Private room

- $\text{price} = 660.3 - 84.5 \text{ minimum_nights} + 12.31 \text{ calculated_host_listings_count} + 0.658 \text{ availability_365}$

Shared room

- $\text{price} = 672 - 11.70 \text{ number_of_reviews} + 288 \text{ reviews_per_month} - 11.55 \text{ calculated_host_listings_count}$

Donde se obtuvo el mejor resultado fue con “Hotel”, obteniendo una R-cuadrada de 0.0467 y un coeficiente de correlación con valor de 0.2161, siendo este el mayor de los 4 elaborados.

Londres

Se realizó un análisis de regresión dentro de cada subconjunto para cada una de las 6 variables mencionadas anteriormente, donde la variable endógena para cada uno de estos fue la de el precio de el lugar de alojamiento. En el caso de “Entire home”, los coeficientes con un valor más cercano a 1, fueron: reviews per month, calculated host listings count y availability 365, respectivamente. Para “Hotel” se sustituye number of reviews por reviews per month, y para calculated host listings count y availability 365. En cuanto a “Private Room” los valores más altos fueron las mismas variables que nuestro primer tipo de alojamiento. Y por último, para “Shared Room”, los coeficientes más altos fueron: minimum nights, calculated host listing counts y availability 365.

Como fue explicado anteriormente, el coeficiente de correlación nos indica la fortaleza y la dirección de la relación entre variables. Al igual que en México, es posible observar que todas las relaciones son positivas, indicando que a medida que aumenta o disminuye un factor el otro también aumentará o disminuirá. De igual manera, se encontraron coeficientes de correlación muy débiles, indicando que las variables seleccionadas no necesariamente tienen un impacto significativo en el precio. Se seleccionaron las 3 variables con el coeficiente de correlación más alto para generar una regresión múltiple. A través de la tabla es posible observar que para todos los tipos de habitación se seleccionaron “calculated_host_listings_count” y “availability_365”. A pesar de que los coeficientes de correlación no son muy altos, es posible inferir que de las variables con las que se está trabajando son las que mejor pueden describir la variable de precio.

Coeficientes de correlación							
LONDON	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	Múltiple
Entire home	0.04123105626	0.03872983346	0.1178982612	0.1946792233	0.2431049156	0.02	0.310322413
Hotel	0.1459451952	0.1846618531	0.1749285568	0.178325545	0.2090454496	0.068556546	0.3028200786
Private room	0.0469041576	0.05099019514	0.09110433579	0.2744084547	0.2338803113	0.02	0.3529872519
Shared room	0.1166190379	0.09539392014	0.04	0.1048808848	0.1236931688	0.1029563014	0.224053565

Coeficientes de determinación							
LONDON	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	Múltiple
Entire home	0.0017	0.0015	0.0139	0.0379	0.0591	0.0004	0.0963
Hotel	0.0213	0.0341	0.0306	0.0318	0.0437	0.0047	0.0917
Private room	0.0022	0.0026	0.0083	0.0753	0.0547	0.0004	0.1246
Shared room	0.0136	0.0091	0.0016	0.011	0.0153	0.0106	0.0502

Al generar estos modelos se obtuvieron coeficientes relativamente altos, incluso uno de 0.35 en el caso de la habitación privada. En todos los tipos de habitaciones, tanto el coeficiente de correlación como el coeficiente de determinación más alto se obtuvo a través del modelo de regresión múltiple. Por ende, se seleccionaron como los modelos matemáticos que mejor describen la variable de precio en todos los casos. Las ecuaciones se pueden observar a continuación;

Entire home

- price = 139.41 + 19.18 reviews_per_month + 0.7367 calculated_host_listings_count + 0.27742 availability_365

Hotel room

- price = 181.7 - 1.306 number_of_reviews + 0.403 calculated_host_listings_count + 0.2405 availability_365

Private room

- price = 48.369 + 5.13 reviews_per_month + 0.8255 calculated_host_listings_count + 0.14274 availability_365

Shared room

- price = 47.59 + 7.61 minimum_nights - 1.516 calculated_host_listings_count + 0.1016 availability_365

Donde se obtuvo el mejor resultado fue con "Private Room", obteniendo una R-cuadrada de 0.1246 y un coeficiente de correlación con valor de 0.3529, siendo este el mayor de los 4 elaborados.

Singapur

Por último, se llevó a cabo el mismo análisis de correlación para Singapur. Al igual que para los países anteriores, se dividió la base de datos por tipo de habitación y se analizó la correlación entre 6 variables y el precio como variable endógena. En el caso de Singapur también se obtuvieron coeficientes positivos, indicando una relación positiva entre las variables. Es decir, el precio tendrá el mismo cambio en dirección que las variables seleccionadas. En este caso no se encontró un patrón tan definido como en el

caso de los otros países. Es decir, no había una variable específica que fuera la más alta en correlación con el precio para todos los tipos de propiedades.

Sin embargo, fue posible encontrar coeficientes de correlación mucho más altos que en los otros casos. Por ejemplo, en el caso del hogar completo se encontró un coeficiente de correlación entre precio y “calculated_host_listings_count” de 0.369, lo cual indica una correlación moderada. Para el hotel la correlación más alta encontrada también fue con “calculated_host_listings_count”, para habitaciones privadas fue “minimum_nights” y para habitaciones compartidas fue “number_of_reviews”. Utilizando las 3 variables con el coeficiente de correlación más alto se generó una regresión múltiple, al igual que en los países ya trabajados. Al obtener estos modelos se obtuvieron coeficientes de correlación y determinación aún más altos. En todos los casos se obtuvo que el modelo matemático que mejor describe la variable de precio es el de regresión múltiple. En el caso del hogar completo incluso se obtuvo un coeficiente de correlación de 0.43, el más alto de todos los obtenidos a través de este análisis.

Coeficientes de correlación							
SINGAPUR	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	Múltiple
Entire home	0.2673948391	0.2317326045	0.05291502622	0.3697296309	0.1723368794	0.2844292531	0.4316248371
Hotel	0.01732050808	0.2317326045	0.02	0.2340939982	0.1104536102	0.06633249581	0.3778888726
Private room	0.223383079	0.08717797887	0.137113092	0.15	0	0.04	0.2771281292
Shared room	0.04795831523	0.2437211521	0.2227105745	0.04358898944	0.1307669683	0.06164414003	0.3670149861

Coeficientes de determinación							
SINGAPUR	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	Múltiple
Entire home	0.0715	0.0537	0.0028	0.1367	0.0297	0.0809	0.1863
Hotel	0.0003	0.0537	0.0004	0.0548	0.0122	0.0044	0.1428
Private room	0.0499	0.0076	0.0188	0.0225	0	0.0016	0.0768
Shared room	0.0023	0.0594	0.0496	0.0019	0.0171	0.0038	0.1347

Se consideraron los 3 coeficientes más altos de cada categoría para así poder elaborar la ecuación de regresión.

Entire home

- price = 162.82 + 0.2563 minimum_nights + 0.6683 calculated_host_listings_count - 3.301 number_of_reviews_ltm

Hotel room

- price = 204.8 - 1.925 number_of_reviews - 6.87 calculated_host_listings_count + 0.1810 availability_365

Private room

- price = 180.6 - 0.7285 minimum_nights + 26.42 reviews_per_month - 0.554 calculated_host_listings_count

Shared room

- price = 77.5 - 1.226 number_of_reviews + 57.7 reviews_per_month - 0.0644 availability_365

Donde se obtuvo el mejor resultado fue con “Entire Home”, obteniendo una R-cuadrada de 0.1863 y un coeficiente de correlación con valor de 0.4316, siendo este el mayor de los 4 elaborados.

Comportamiento de las ciudades elegidas

Se realizó un análisis comparando las ciudades de México, Singapur y Londres para encontrar factores críticos que se pudieran extender a México. Es decir, si encontrábamos que el mínimo de noches tenía un efecto significativo en el precio a través de los diferentes tipos de habitación tanto en Singapur como en Londres esto podría representar un hallazgo. Con base en lo anterior sería posible explorar a mayor profundidad la variable encontrada y hacer recomendaciones. De cualquier manera, no se encontró ningún patrón significativo que se pudiera extender a todos los países. Asimismo, no se encontraron coeficientes de correlación lo suficientemente altos como para ser considerados una correlación fuerte. Es importante mencionar que el precio en sí es una variable fijada por el anfitrión, por ende, no necesariamente tiene una respuesta a las otras variables y eso podría explicar lo débiles que fueron las correlaciones encontradas.

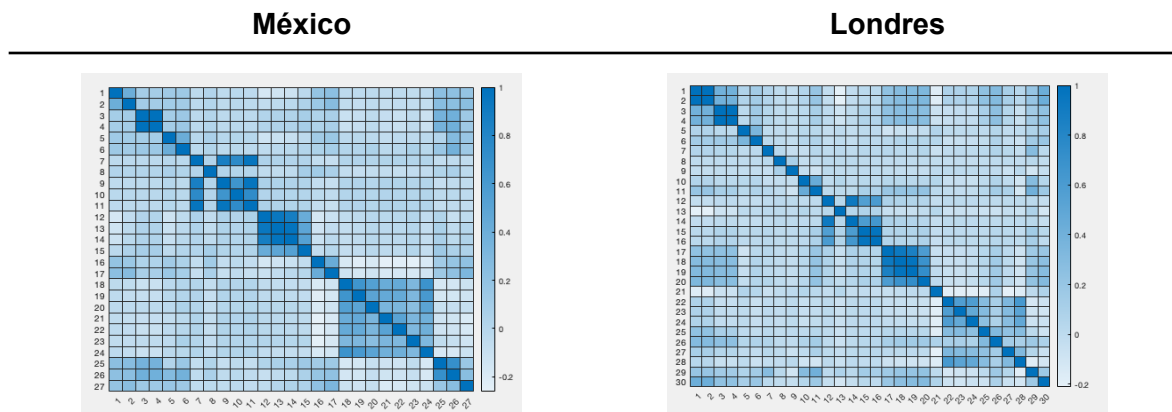
Para los 3 países se utilizaron las 3 variables que tuvieron el coeficiente de correlación más alto con la variable de precio para crear un modelo de regresión múltiple. Al hacer esta selección para cada tipo de habitación se pudo observar un patrón entre las variables. Por ejemplo, para Londres en todos los tipos de habitación las variables de “calculated_host_listings_count” y “availability_365” tuvieron el coeficiente de correlación más alto, pudiendo categorizarlas como los factores críticos de los incluidos en la base de datos. De cualquier manera, el mismo comportamiento no se pudo extender a todos los países. A pesar de esto, sí fue posible visualizar que la variable de “calculated_host_listings_count” fue la variable que mayores coeficientes de correlación tuvo para los diferentes tipos de habitación y diferentes países. Por ende, se podría decir que generalmente la variable que mejor describe el precio es “calculated_host_listings_count”. Por otra parte, la variable con menores coeficientes de correlación encontrados fue la de “number_of_reviews_ltm”, lo que indica que sería la menos descriptiva de las encontradas en la base de datos.

A pesar de no encontrar resultados lo suficientemente significativos como para hacer conclusiones que se puedan extender a todos los países o desarrollar estrategias en base a los resultados, fue posible conocer más acerca de la plataforma alrededor de los países. Por ejemplo, fue posible encontrar que los factores críticos cambian por país, por ende, la cultura sí tiene un efecto en el uso de la plataforma. Esto es de utilidad ya que México tiene que encontrar los factores críticos de éxito de este tipo de plataformas en México en base al conocimiento del mercado, de la industria, de la cultura, y más dentro del país. Es decir, el utilizar estrategias mercadológicas de otros países en México para aumentar el uso de la plataforma no sería de mucha utilidad ya que sí hay un cambio de comportamiento por país.

COMPLEMENTO

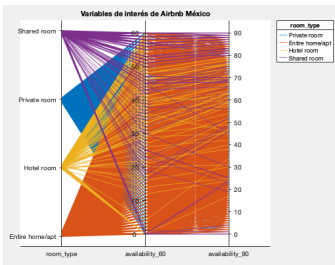
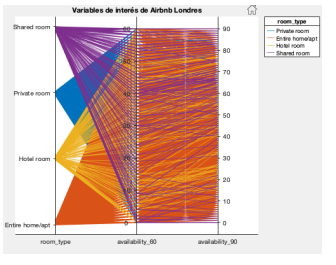
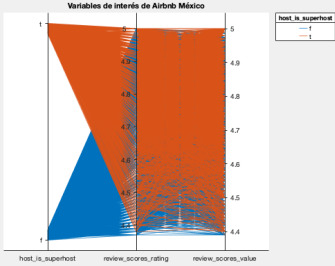
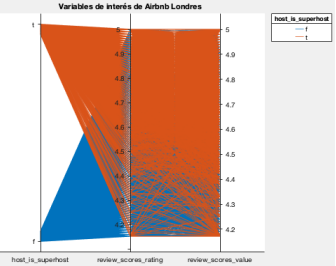
Como fue mencionado anteriormente, una parte del objetivo de este análisis también consiste en comparar plataformas de inteligencia de negocios. Por ende, se utilizó la plataforma de Matlab para continuar con este análisis. En este caso, se trabajó con 2 países; México y Londres. Para ambos países se realizaron acciones de preprocesamiento para los valores nulos y atípicos encontrados. Resultó más tedioso el procedimiento en esta plataforma que utilizando Python. Se llevaron a cabo procesos muy similares para la limpieza de la base de datos de Londres y de México para no afectar los resultados. En un principio, los procedimientos fueron adaptados debido a que las bases de datos eran diferentes desde el inicio del proceso. De cualquier manera, al graficar las diferentes variables pudimos observar que las escalas cambiaban y que por ende, no se podían comparar. Debido a esto, se aplicaron los mismos métodos para ambas bases de datos en la sustitución de valores nulos y valores atípicos.

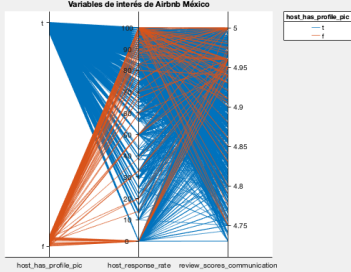
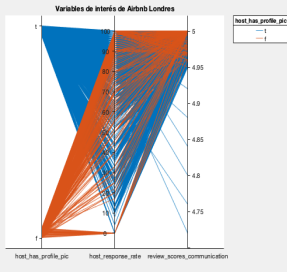
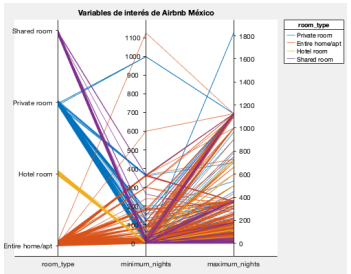
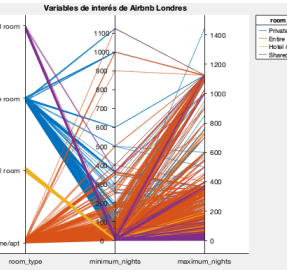
Se llevaron a cabo los análisis de correlación entre todas las variables cuantitativas incluidas en la base de datos de cada país y se obtuvieron los siguientes gráficos.

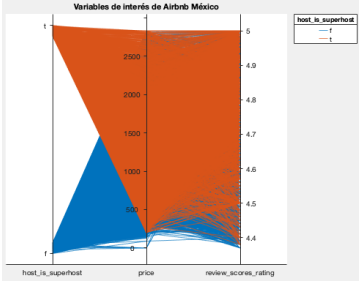
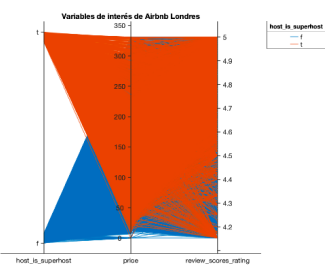


Al visualizar los gráficos en Matlab es posible leer los coeficientes de correlación entre las variables incluidas. Es importante mencionar que a pesar de haber filtrado la base de datos, tomando en cuenta sólo ciertas columnas, algunas resultaron no tener correlación en lo absoluto. Por ende, se creó otro filtro con menos columnas y los gráficos que se pueden observar son resultado de ese filtro. De cualquier manera, al analizarlo más a detalle se pudo observar que las correlaciones más altas eran entre variables que dependen de sí mismas. Por ejemplo, el mínimo y máximo de noches viene como diferentes combinaciones en la base de datos y las correlaciones fuertes se encuentran concentradas en esas combinaciones. Esto se puede interpretar desde el gráfico ya que en la base de datos las variables similares entre sí están juntas. En el gráfico se puede observar que las correlaciones fuertes están concentradas, por ende, se puede observar esa correlación.

Resultó más tedioso trabajar con Matlab en este caso ya que se altera el orden de las variables entonces se tenía que mantener una tabla para no perder el orden de las variables e interpretarlas de manera correcta. Como fue mencionado anteriormente, no se encontraron correlaciones significativas entre variables que no fueran dependientes de sí mismas, como por ejemplo: `minimum_nights` con `minimum_minimum_nights` o combinaciones del estilo para ninguno de los 2 países utilizados en el análisis. De cualquier manera, se realizaron 10 gráficos con las correlaciones más altas encontradas o con variables de interés. Al graficar las 10 gráficas pudimos observar que era posible desarrollar un análisis comparativo a través de ciertas gráficas entre México y Londres.

MÉXICO	LONDRES	ANÁLISIS
		<p>Se puede observar un patrón similar de comportamiento entre los consumidores de ambos países. De cualquier manera, se puede observar que México tiene menor disponibilidad en plazos menores a 60 días mientras que en Londres se encuentra mucho mayor disponibilidad en un rango mayor de días. Esto indica que las estadías cortas en México son comunes, esto podría ser debido a los precios que se ofrecen en la Ciudad de México.</p>
		<p>En este caso también se pudo observar un comportamiento similar entre ambos países. Se puede observar que el hecho de que el host sea un superhost concentra la mayoría de los datos en un <code>review_scores_value</code> más alto. En ambos países se puede observar que si el anfitrión es un superhost la mayoría de los <code>review_scores_rating</code> se concentra en valores por</p>

		<p>arriba de 4.6. En este caso es posible que debido a que esos valores están ubicados por arriba del 4.6 el anfitrión es categorizado como superhost.</p>
		<p>Existe un mayor número de anfitriones en Londres que no cuentan con foto en su perfil a comparación de México. Pero los mismos anfitriones londinenses cuentan con una buena calificación en satisfacer a sus clientes en cuestión de comunicación, este dato lo podemos comprobar tanto en el rate de respuesta como en el score de comunicación. Se puede observar que los anfitriones de Londres tienen una mejor comunicación ya que la distribución de los puntajes de las reseñas se encuentran mayormente concentrados en el 5. De cualquier manera, en México parecen ser resultados más precisos entre el ratio de respuesta del anfitrión y los puntajes que recibieron.</p>
		<p>En este caso también se puede observar que la distribución de los datos es similar en ambos países. Por ejemplo, se observa que para el hogar completo las noches mínimas van de un rango de 0-100 aproximadamente y de máximas de 0-1200. Por otra parte, para los hoteles se observa que el rango de noches mínimas y máximas</p>

		es menor. Por ende, ambos países demuestran tener políticas de estadía similares en cuanto a rango de tiempo.
		<p>Tanto del lado de México como de Londres se puede notar una calificación positiva hacia los dueños de los lugares que rentan los clientes en airbnb. Incluso en las calificaciones negativas, con la diferencia de que una pequeña cantidad de la muestra en México se encuentra en un punto más bajo del gráfico. Se interpreta que los dueños de localidades en Airbnb tienen un buen servicio y eso les hace tener una buena reputación, aun cuando los precios puedan ser altos o bajos, esto no influye en la satisfacción de los clientes.</p>

Tanto las gráficas incluidas en el análisis como las no incluidas se pueden visualizar en Matlab. De cualquier manera, consideramos que estas 5 gráficas son las que mejor permitirían comparar el comportamiento de las variables en los países ya que algunas eran más específicas e incluían la ubicación. A través de este análisis fue posible encontrar factores que tienen el mismo patrón a través de todos los países, como la correlación entre `calculated_host_listings` y el precio. Además fue posible observar prácticas comunes como el rango del mínimo y máximo de noches y visualizar que en dos países en diferentes continentes se comportan de una manera similar. Por otra parte, fue interesante rescatar factores en los que los países difieren. Por ejemplo, el hecho de que en México más anfitriones tengan foto de perfil que en Londres puede ser a causa del tema de seguridad en México.

A pesar de no haber encontrado causalidades o factores sumamente relevantes a través de este análisis, fue posible analizar el desempeño de la plataforma en diferentes culturas, idiomas, y países. De la misma manera, fue posible evaluar el desempeño de las plataformas en diferentes análisis para poder hacer una recomendación de uso más específica en el futuro considerando factores como la curva de aprendizaje, la visualización, la eficacia de las plataformas, y más.

Bibliografía:

Get the Data. (2022). Recuperado Octubre 11, 2022, de Insideairbnb.com website:

<http://insideairbnb.com/get-the-data/>