Actividad 5 (Regresión Lineal Simple)

Nombre del Equipo

CAPT

Miembros

Danna Paola Arciniega Zúñiga | A01731987

Victoria Eugenia Téllez Castillo | A01732258

Mitzi Castelán Chávez | A01731147

Elizabeth Pérez García | A01366971

Fecha

06 de Octubre de 2023

Analítica de datos y herramientas de inteligencia artificial II

A lo largo de esta actividad se nos proporcionan dos bases de datos “California USA” y “México”, con ellas se estarán realizando acciones de procesamiento necesarias para poder eliminar datos innecesarios como lo son nulos y outliers, de igual forma se creará un modelo matemático donde se describa de la mejor manera y el número de reseñas para cada tipo de alojamiento, utilizando la variable con mayor correlación, dentro del documento también se mostrará una tabla de todos los coeficientes de determinación y correlación obtenidos para cada tipo de habitación elegido.

DF México

Primeramente, comenzaremos con la base de datos “México”, para ella se comenzó importando las librerías necesarias, en continuación importamos nuestra base de datos, la leemos y pedimos información principal para saber el estado de la misma.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import files
files.upload()
df=pd.read_csv('DF_Mexico.csv')
df.info
```

Figura 1. Lectura de datos

Con esto en cuenta seleccionamos las columnas que vamos a utilizar, utilizamos la función de loc la cual devuelve el elemento que se encuentra en la columna de con nombre columna del DataFrame, con ello en continuación identificamos valores nulos para después reemplazarlos para poder obtener 0 para cada columna, quedando de la siguiente manera:

```
#IDENTIFICAR VALORES NULOS
valores_nulos=df1.isnull().sum()
valores_nulos
```

```
host_acceptance_rate          2043
host_response_rate            3091
review_scores_location        3653
review_scores_cleanliness     3651
price                            0
availability_365                 0
number_of_reviews                0
reviews_per_month             3596
review_scores_communication   3652
dtype: int64
```

```
#reemplazamos los valores nulos
df1=df1.fillna(method="bfill")
df2=df1.fillna(method="ffill")


valores_nulos1=df2.isnull().sum()
valores_nulos1
```

```
host_acceptance_rate          0
host_response_rate            0
review_scores_location        0
review_scores_cleanliness     0
price                         0
availability_365              0
number_of_reviews             0
reviews_per_month             0
review_scores_communication   0
dtype: int64
```

Figura 2. Identificación y reemplazo de valores nulos

El siguiente paso fue añadir la variable llamada ‘room_type’, la cuál es la encargada de ser el filtro que se va utilizar en ambos DataFrames, tanto en el de México como en el de USA. Al decir filtro, nos referimos a lo siguiente; la manera en la que vamos a separar nuestro conjunto de datos es mediante las diferentes opciones de room_type (son 4 en total, Entire home/apt, Private room, Hotel room y Shared room) en donde generamos 2 Data Frames de cada uno de los originales, resultando un total de 4 Dataframes, dos de México con diferentes tipos de habitaciones, y dos de USA con las mismas características. Para seleccionar de mejor manera cuál de los 4 filtros es el

más eficiente y completo, decidimos basarnos en el total de columnas de cada uno, al finalizar, se decidió usar el filtro de Entire home/apt y Private Room para continuar con su limpieza de valores atípicos y regresiones. Recordamos, que este proceso lo realizamos para **ambas** bases de datos.

California USA

Continuaremos con la base de datos "California USA", para ella al igual que la pasada, se comenzó importando las librerías necesarias, como segundo punto importamos nuestra base de datos y la leemos y aquí mismo seleccionamos las columnas a utilizar para poder visualizar los datos nulos y outliers.

```
#Carga desde un archivo .xlsx sin indice
USA = pd.read_csv('California_EUA.csv')
USA
```

```
USA_1 = USA.loc [: , ["host_acceptance_rate","host_response_rate", "review_scores_location",
                      "review_scores_cleanliness", "price", "availability_365",
                       "number_of_reviews", "reviews_per_month", "review_scores_communication"]]
USA_1
```

Figura 3 . Lectura de datos y selección de columnas

En continuación corroboramos los datos nulos, los reemplazamos y volvemos a verificar que se hayan llenado correctamente, esto quedando de la siguiente manera.

```
#Corroborar valores nulos
valores_nulos = USA_1.isnull().sum()
valores_nulos
```

```
host_acceptance_rate          801
host_response_rate           1019
review_scores_location       1347
review_scores_cleanliness    1346
price                           0
availability_365                0
number_of_reviews               0
reviews_per_month            1325
review_scores_communication  1346
dtype: int64
```

```
#Remplazar valores nulos del dataframe con bfill
USA_1_CLEAN = USA_1.fillna(method='bfill')
USA_1_CLEAN = USA_1_CLEAN .fillna(method='ffill')
```

```
#Corroborar valores nulos
valores_nulos = USA_1_CLEAN.isnull().sum()
valores_nulos
```

```
host_acceptance_rate         0
host_response_rate           0
review_scores_location       0
review_scores_cleanliness    0
price                        0
availability_365             0
number_of_reviews            0
reviews_per_month            0
review_scores_communication  0
dtype: int64
```

Figura 4. Identificación de valores nulos

La limpieza de valores atípicos se hicieron de la misma manera en todos los procesos, se hizo uso del procedimiento de desviación estándar para la eliminación de estos valores, consideramos este procedimiento como el más óptimo dentro del proceso de limpieza ya que consideramos todos los valores como atípicos cuando se encuentran alejados a 3 desviaciones típicas de la media de los valores. Usamos las condiciones necesarias, eliminamos estos valores y ahora que se categorizan como nulos, los reemplazamos con la media.

Escala de Correlación

El coeficiente puede variar de -1 a 1, donde el signo indica la dirección de la correlación y el valor numérico, la magnitud de la correlación. En este contexto se resumen algunos criterios de interpretación:

-1	Correlación negativa perfecta
-0,90	Correlación negativa muy fuerte
-0,75	Correlación negativa considerable
-0,50	Correlación negativa media
-0,10	Correlación negativa débil
0	Correlación nula

+1	Correlación positiva perfecta
+0,90	Correlación positiva muy fuerte
+0,75	Correlación positiva considerable
+0,50	Correlación positiva media
+0,10	Correlación positiva débil
0	Correlación nula

Comparación

MÉXICO ENTIRE HOME/APT

Columnas	Modelo Matemático	Coeficiente Determinación	Coeficiente Correlación
x=host_response_rate y=host_acceptance_rate	y=0.467x + 50.2289	0.0753	0.274
x=price y='host_acceptance_rate	y=-4.668x + 96.271	1.24993	0.00112
x=number_of_reviews y=host_acceptance_rate	y= 0.025x + 95.379	0.01726	0.131
x=review_scores_cleanliness y=review_scores_location	y=0.218x + 3.826	0.07817	0.2796
x=number_of_reviews y=availability_365	y= -0.1553x + 246.317355	0.00347	0.05894
x=reviews_per_month y=review_scores_communication	y= 0.00225x + 4.867	0.0002795	0.016718

MÉXICO PRIVATE ROOM

Columnas	Modelo Matemático	Coeficiente Determinación	Coeficiente Correlación
x=host_response_rate y=host_acceptance_rate	y=0.3838x + 52.4581	0.1254	0.35415
x=price y='host_acceptance_rate	y= -4.02004x + 87.455222	1.424	0.00377
x=number_of_reviews y=host_acceptance_rate	y= 0.05639x + 85.98978	0.0162	0.1275
x=review_scores_cleanliness y=review_scores_location	y= 0.6511x + 1.74012	0.5375	0.733
x=number_of_reviews y=availability_365	y= -0.093645x + 255.47449	0.00185	0.043
x=reviews_per_month y=review_scores_communication	y=0.010820x + 4.7750508	0.0020	0.04483

USA ENTIRE HOME/APT

Columnas	Modelo Matemático	Coeficiente Determinación	Coeficiente Correlación
x=host_response_rate y=host_acceptance_rate	y = 0.11236909 + 79.5696807714275	0.01355	0.116444
x=price y='host_acceptance_rate	y= -0.00040132 + 90.17235984921105	0.001580	0.039759
x=number_of_reviews y=host_acceptance_rate	y = 0.02972475 + 88.32305311819619	0.02259	0.150309

x=review_scores_cleanliness y=review_scores_location	y= 0.51876403 + 2.3075065977577927	0.31349	0.5599
x=number_of_reviews y=availability_365	y = -0.00252966 + 177.4508942427995	3.65396	0.001911
x=reviews_per_month y=review_scores_communication	y = 0.01519389 + 4.795589869198288	0.0064077	0.08004

USA PRIVATE ROOM

Columnas	Modelo Matemático	Coeficiente Determinación	Coeficiente Correlación
x=host_response_rate y=host_acceptance_rate	y =0.08208067 + 85.0875	0.00265	0.0515
x=price y='host_acceptance_rate	y = -0.00055144 + 93.138	0.0001255	0.0112
x=number_of_reviews y=host_acceptance_rate	y = 0.02946289 + 91.74646	0.0319	0.1787
x=review_scores_cleanliness y=review_scores_location	y = 0.33815943 + 3.186	0.313492	0.55990
x=number_of_reviews y=availability_365	y= -0.08288418 + 180.9834	0.001989	0.0446
x=reviews_per_month y=review_scores_communication	y = 0.00963526 + 4.847	0.00378	0.0614

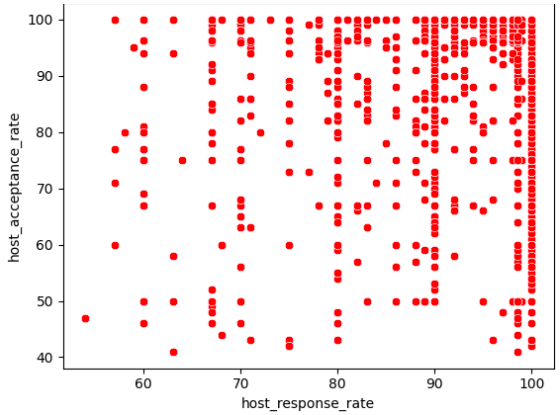
Primera regresión HOME/APT

Realizamos la comparación de columnas, donde identificamos la correlación, se sabe que en cuanto más cerca de +1, más alta es su asociación, un valor exacto de +1 indicaría una relación lineal positiva perfecta, Para las siguientes gráficas se mostrarán en forma de dispersión, su variabilidad se va a referir a la distancia entre los puntos de datos y la línea de tendencia, los puntos de datos que están más alejados de la línea de tendencia se consideran más variables.

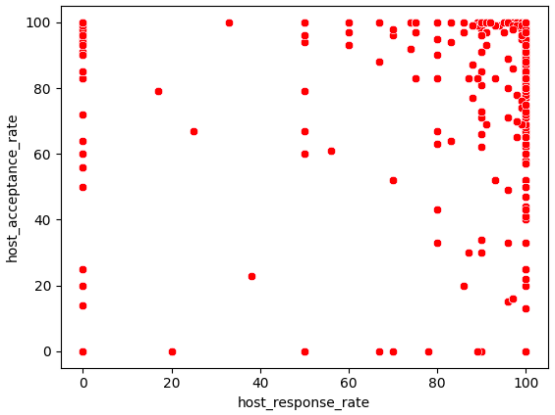
Una gráfica de dispersión con mayor variabilidad indica que hay una mayor dispersión de los datos, significa que los puntos de datos están más dispersos y, por lo tanto, es más probable que la

relación entre las dos variables sea menos fuerte y predecible, a continuación se muestran las 6 gráficas referente a los previos valores de cada una.

México



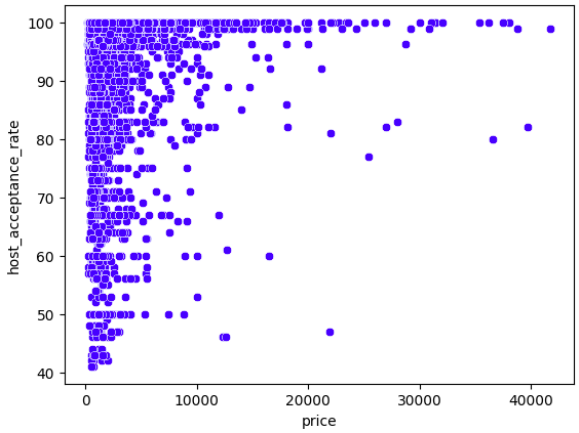
USA



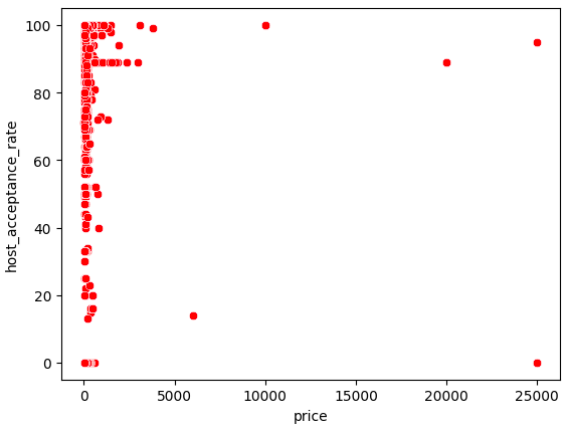
Variable independiente—> x=host_response_rate
Variable dependiente—> y=host_acceptance_rate

Para las previas imágenes se puede tener una comparación y se pueden observar diferencias notables entre las dos gráficas, la gráfica de México muestra una mayor dispersión de los datos que la gráfica de Estados Unidos, esto significa que hay una mayor variabilidad en la tasa de respuesta de anfitriones para cada nivel de tasa de aceptación de anfitriones.

México



USA

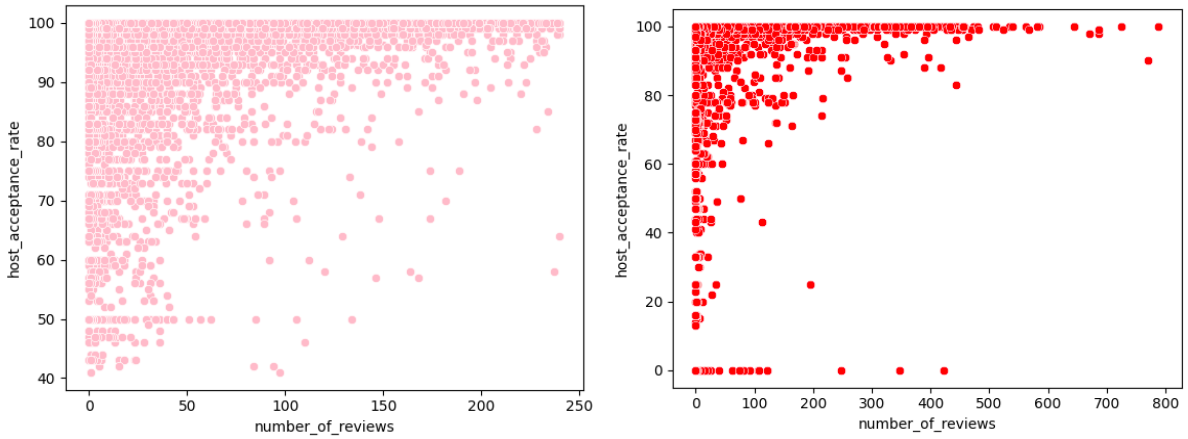


Variable independiente—> x=price
Variable dependiente—> y=host_acceptance_rate

Para las dos gráficas, muestran que existe una correlación positiva entre el precio y la tasa de aceptación de anfitriones, esto significa que los anfitriones que cobran precios más altos tienden a tener tasas de aceptación de anfitriones más altas, en la gráfica de México se observa una mayor dispersión de los datos que la gráfica de Estados Unidos, esto significa que hay una mayor variabilidad en la tasa de aceptación de anfitriones para cada nivel de precio.

México

USA

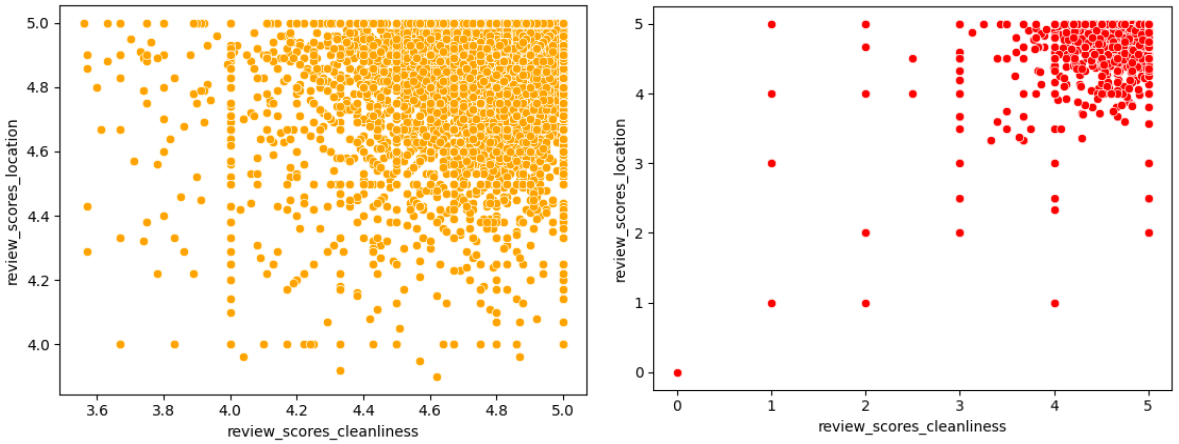


Variable independiente—> x=number_of_reviews
Variable dependiente—>y=host_acceptance_rate

La gráfica de México muestra una mayor dispersión de los datos que la gráfica de Estados Unidos,significa que hay una mayor variabilidad en la tasa de aceptación de anfitriones para cada nivel de número de reseñas, en México se observa que los anfitriones con 100 reseñas tienen tasas de aceptación de anfitriones que van desde 20% hasta 100%. Comparado con Estados Unidos, los anfitriones con 100 reseñas tienen tasas de aceptación de anfitriones que van desde 80% hasta 100%, esto indica que hay menos variabilidad en la tasa de aceptación de anfitriones en Estados Unidos.

México

USA

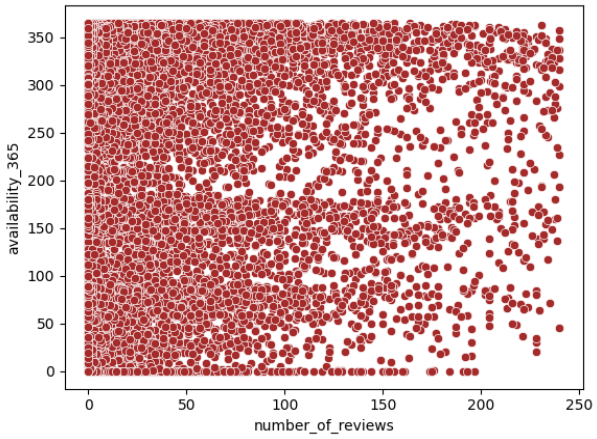


Variable independiente—> x=review_scores_cleanliness
Variable dependiente—>y=review_scores_location

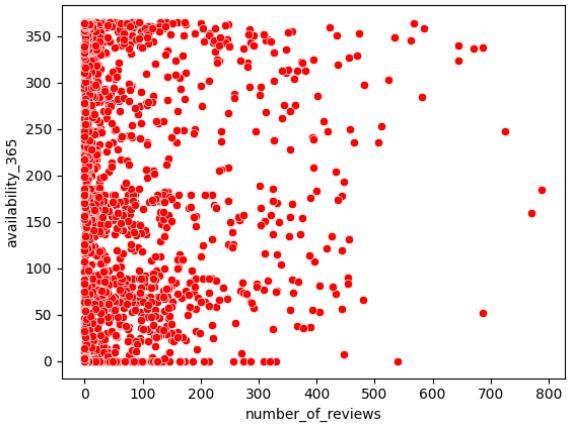
Las dos gráficas muestran una correlación positiva entre las puntuaciones de limpieza de las reseñas y las puntuaciones de ubicación de las reseñas, significa que los alojamientos con puntuaciones de limpieza más altas tienden a tener puntuaciones de ubicación más altas, la gráfica de México muestra una mayor dispersión de los datos que la gráfica de Estados Unidos. Esto significa que hay una mayor variabilidad en las puntuaciones de ubicación de las reseñas en

México. la gráfica de México en si fue la mejor con respecto al resultado del coeficiente de determinación quedando con un valor de 0.78, de igual forma en ambos casos, el coeficiente de correlación es moderado, pero el modelo de los Estados Unidos tiene un coeficiente de correlación más alto (0.5599) en comparación con el modelo de México (0.2796), lo que indica una relación más fuerte entre las variables en los Estados Unidos.

México



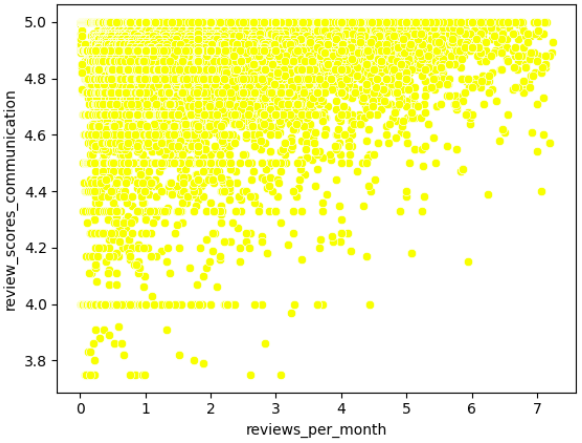
USA



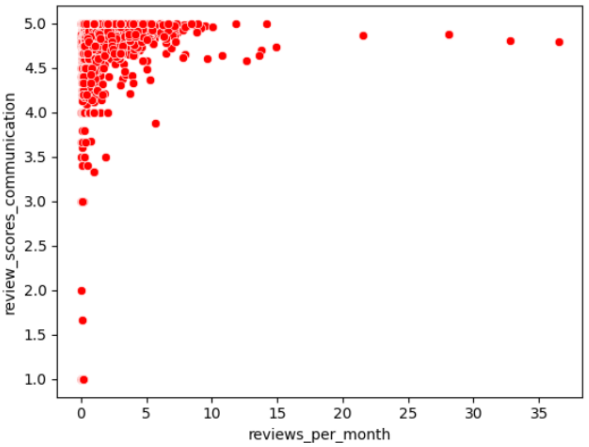
Variable independiente—> x=number_of_reviews
Variable dependiente—>y=availability_365

Las dos gráficas muestran una correlación negativa entre el número de reseñas y la disponibilidad de los alojamientos, significa que los alojamientos con más reseñas tienden a tener una menor disponibilidad, La gráfica de México muestra una mayor dispersión de los datos que la gráfica de Estados Unidos, esto quiere decir que hay una mayor variabilidad en la disponibilidad de los alojamientos en México, se observa y se puede predecir que el coeficiente de determinación para usa es de 3.65, por lo que es significativamente alto en comparación con México

México



USA

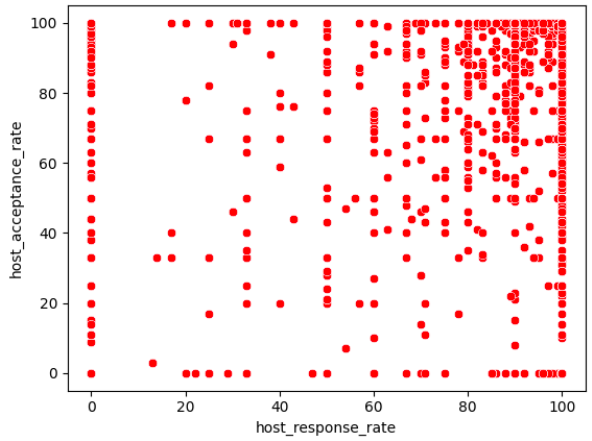


Variable independiente—> x=review_per_month
Variable dependiente—>y=review_scores_communication

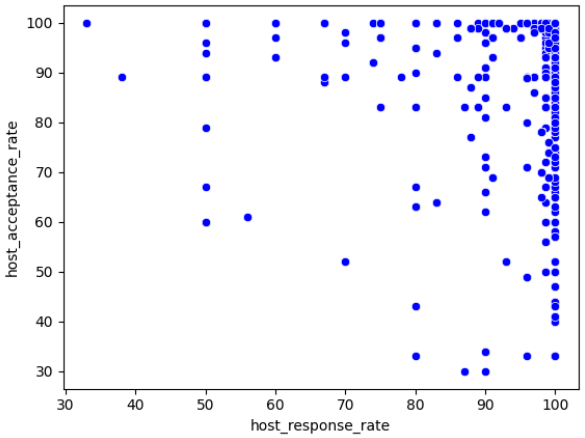
Las dos gráficas muestran una correlación positiva entre el número de reseñas por mes y las puntuaciones de comunicación de las reseñas, se refiere a esto ya que los alojamientos con más reseñas por mes tienden a tener puntuaciones de comunicación más altas, la gráfica de México muestra una mayor dispersión de los datos que la gráfica de Estados Unidos, quiere decir que hay una mayor variabilidad en las puntuaciones de comunicación de las reseñas en México, el coeficiente de determinación para Estados Unidos (0.0064077) es más alto que el de México (0.0002795), lo que indica que el modelo para Estados Unidos explica una mayor proporción de la variabilidad en los datos en comparación con el modelo de México, para los dos gráficos el coeficiente de correlación es bajo, pero el modelo de Estados Unidos tiene un coeficiente de correlación más alto (0.08004) en comparación con el modelo de México (0.016718), lo que indica una relación un poco más fuerte entre las variables en Estados Unidos.

Segunda regresión Private Room

México



USA

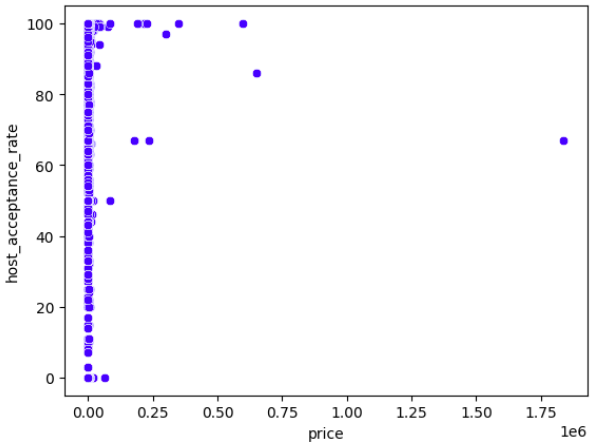


Variable independiente—> x=host_response_rate
Variable dependiente—> y=host_acceptance_rate

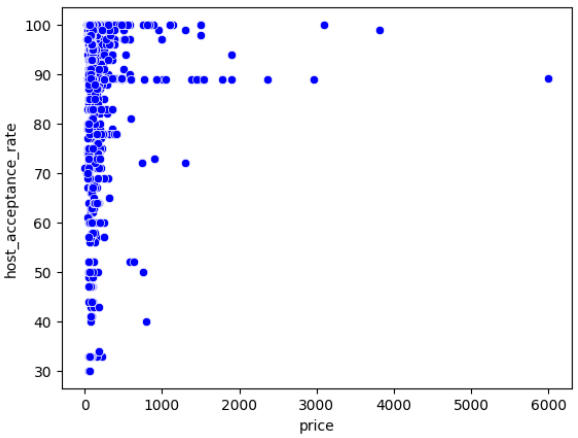
Comenzamos con la segunda regresión tomando en cuenta private room,para estas dos gráficas podemos observar que en la gráfica de México existe mayor dispersión y mayor variabilidad en comparación con la de usa, ambas muestran una correlación positiva entre la tasa de respuesta del anfitrión y la tasa de aceptación del anfitrión, esto quiere decir que los anfitriones que responden más rápido a las solicitudes de los huéspedes tienden a tener una tasa de aceptación más alta, el coeficiente de determinación para Estados Unidos es muy bajo (0.00265), lo que indica que el modelo explica una cantidad muy pequeña de la variabilidad en los datos,comparado con México es más alto (0.1254), se puede decir que este modelo explica una proporción mayor de la variabilidad en los datos, para ambos casos el coeficiente de correlación es bajo, pero el modelo

de México tiene un coeficiente de correlación más alto (0.35415) en comparación con Estados Unidos (0.0515), lo que indica una relación más fuerte entre las variables en México.

México



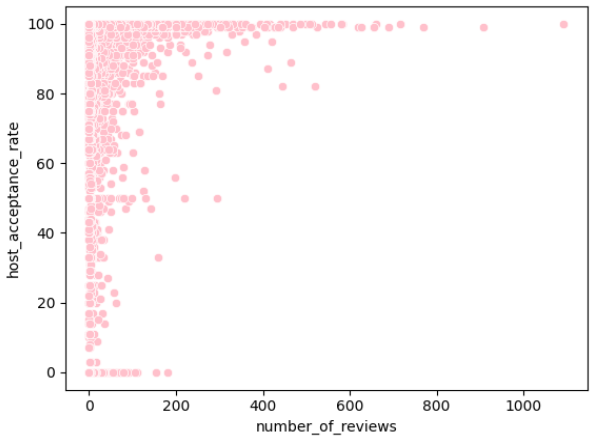
USA



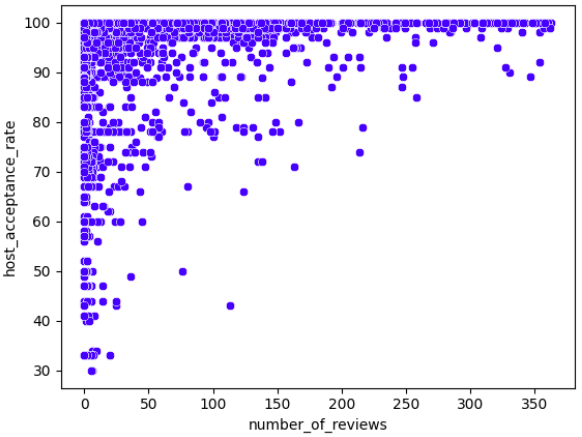
Variable independiente—> x=price
Variable dependiente—> y=host_acceptance_rate

La gráfica de México muestra una mayor dispersión de los datos que la gráfica de Estados Unidos. Esto significa que hay una mayor variabilidad en la tasa de aceptación de anfitriones en México, para la gráfica de Estados Unidos, el modelo matemático es de tipo lineal, lo que indica que existe una relación lineal entre las dos variables, su coeficiente de determinación es muy bajo, lo que indica que el modelo no explica una gran parte de la variación en los datos, también el coeficiente de correlación es bajo, lo que indica que la relación entre las dos variables es débil, el modelo para México se ajusta mucho mejor a los datos en comparación con el modelo de usa, sin embargo, en ambos casos, la relación entre las variables es débil, lo que puede indicar que otros factores no considerados en los modelos pueden estar influyendo en los datos.

México



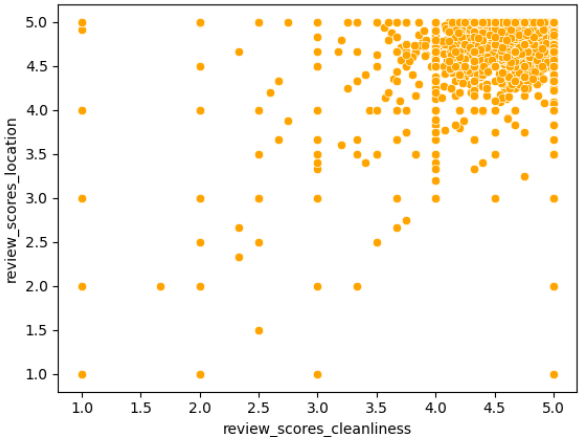
USA



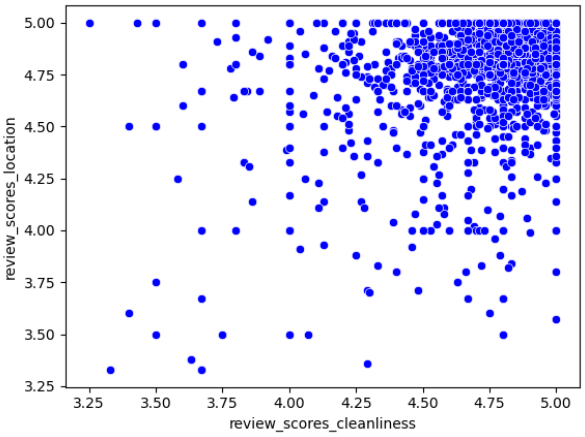
Variable independiente—> x=number_of_reviews
Variable dependiente—>y='host_acceptance_rate

La gráfica muestra que hay una mayor variabilidad en los datos de México que en los datos de Estados Unidos, los modelos matemáticos para ambos países son líneas rectas, lo que indica una relación lineal entre las variables x e y en ambos casos, el coeficiente de determinación es una medida de cuánto de la variabilidad en los datos se explica por el modelo, en este caso, para Estados Unidos es 0.0319, mientras que para México es 0.0162, significa que en ambos casos, los modelos explican una cantidad relativamente baja de la variabilidad en los datos, pero el modelo para Estados Unidos explica un poco más en comparación con el modelo para México, En ambos casos, el coeficiente de correlación es bajo, lo que sugiere que la relación entre las variables es débil, ligeramente Estados Unidos lo tiene un poco más elevado lo que significa que tiene una relación ligeramente más fuerte.

México



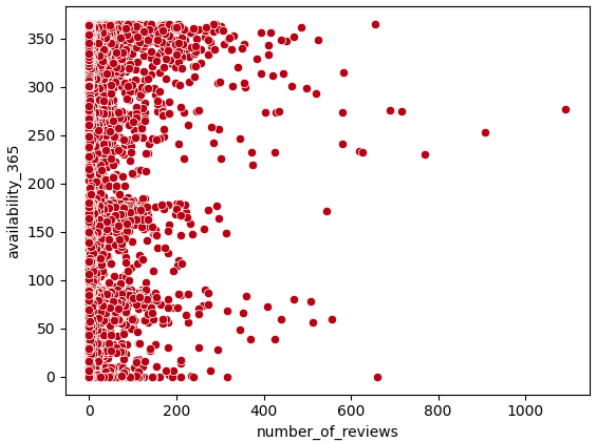
USA



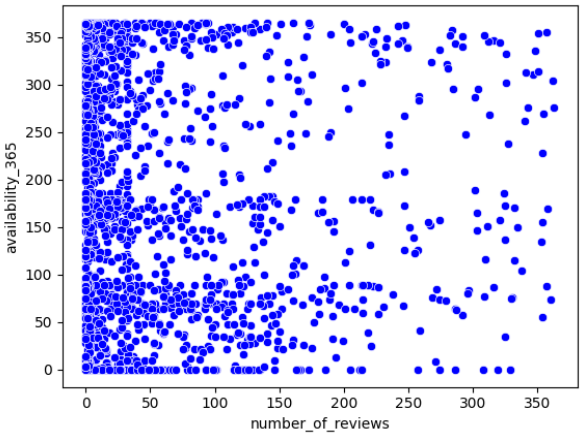
Variable independiente—> x=review_scores_cleanliness
Variable dependiente—>y=review_scores_location

Los gráficos de dispersión proporcionados muestra que existe una correlación positiva entre las variables ya que compara las puntuaciones de limpieza de los alojamientos que tienden a estar correlacionadas con las puntuaciones de ubicación de los alojamientos, comparando los valores obtenidos con su modelo matemático se observa que para ambos es una línea recta, indica una relación lineal entre las variables, por otra parte para los coeficientes de determinación indica que el modelo en México tiene mayor proporción de la variabilidad en los datos en comparación con usa, para el coeficiente de correlación, para ambos modelos es moderado, pero para México su reclutado es mayor, ya que fue de (0.833) en comparación que usa (0.56) esto indica que hay una relación más fuerte entre las variables de México.

México



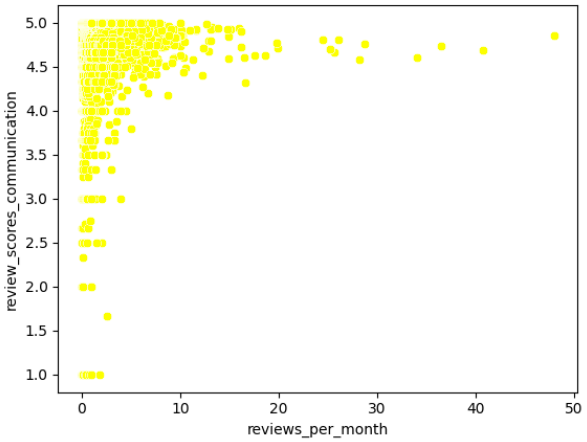
USA



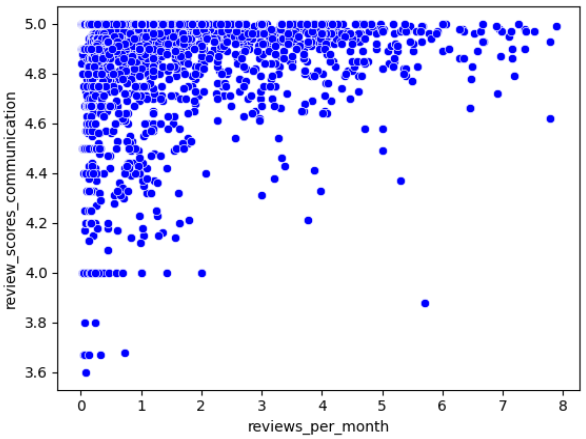
Variable independiente—> x=number_of_reviews
Variable dependiente—>y=availability_365

La gráfica de México muestra una mayor dispersión de los datos que la gráfica de Estados Unidos. Esto significa que hay una mayor variabilidad en la disponibilidad de los alojamientos en México, Los modelos matemáticos para las dos gráficas son similares, pero el modelo para México tiene un coeficiente de determinación ligeramente menor que el modelo para Estados Unidos, significa que el modelo para México explica una menor parte de la variación en los datos, los coeficientes de correlación para las dos gráficas son similares, lo que indica que la relación entre las dos variables es débil en ambos casos.

México



USA



Variable independiente—> x=review_per_month
Variable dependiente—>y=review_scores_communication

Las dos gráficas muestran una correlación positiva entre el número de reseñas por mes y las puntuaciones de comunicación de las reseñas, significa que los alojamientos con más reseñas por mes tienden a tener puntuaciones de comunicación más altas, tanto el coeficiente de determinación para los Estados Unidos (0.00378) como el de México (0.0020) son muy bajos, lo que indica que los modelos explican una cantidad muy pequeña de la variabilidad en los datos, en

ambos casos, el coeficiente de correlación es muy bajo, lo que sugiere que la relación entre el número de revisiones por mes y la puntuación de comunicación en las revisiones es extremadamente débil en ambos países.

Conclusión

Las gráficas de dispersión son importantes en el análisis de datos ya que permiten visualizar de manera efectiva la relación entre dos variables, lo que ayuda a comprender mejor los patrones, tendencias y posibles correlaciones en los datos.

Al analizar las gráficas de dispersión, se puede evaluar la dirección y la fuerza de la correlación entre las dos variables, una correlación positiva indica que a medida que una variable aumenta, la otra también lo hace, mientras que una correlación negativa indica lo contrario, un coeficiente de correlación cercano a 1 o -1 indica una correlación fuerte, mientras que valores cercanos a 0 indican una correlación débil, para los ejemplos anteriores, se compararon gráficas de dispersión y modelos matemáticos para México y Estados Unidos en diferentes contextos, como puntuaciones de revisión, tasas de respuesta de los anfitriones y tasas de aceptación de los anfitriones. En general, las gráficas de dispersión y las métricas (modelo matemático, coeficiente de determinación y coeficiente de correlación) se utilizaron para evaluar la relación entre las variables en cada caso. Se encontró que en algunos casos, como las puntuaciones de revisión, había relaciones más fuertes en un país que en el otro, mientras que en otros casos, como el número de revisiones y la disponibilidad, las relaciones eran débiles en ambos países.