REPORTE COMPARATIVO ACTIVIDAD 1

MATERIA: ANALÍTICA DE DATOS E INTELIGENCIA ARTIFICIAL 2

Luis Javier Gutiérrez Márquez A01734254

OBJETIVO:

Realizar un análisis de las bases de datos proporcionadas de Airbnb desde la fuente **http://insideairbnb.com/get-the-data/** en específico de la ciudad de México y para fines de estudio se optó por realizar el análisis de las ciudades de Chicago y Nueva York para después realizar comparaciones.

El objetivo del análisis es determinar si las variables: host_acceptance_rate, review_scores_rating, Price, review_scores_cleanliness, availability_365 y review_scores_communication. Tienen una correlación con el número de reseñas para 3 diferentes tipos de cuartos (Private_Room, Shared_Room y Hotel_Room).

Para realizar todo el análisis de los datos proporcionados se utilizarán las herramientas de Google colab, Visual Studio Code, Excel, Github y Git.

PLANTEAMIENTO:

Para poder realizar el análisis debemos de revisar que se cuenta con una base de datos limpia por lo tanto se deberá realizar un estudio de datos nulos y outliers para cada una de las tablas de datos. Una vez sustituyendo valores nulos y la base de datos limpia para trabajar se analizará la correlación que existe en cada tipo de habitación respecto a las variables descritas anteriormente.

Una vez visualizado las correlaciones se optará por escoger 1 sola variable para crear el modelo matemático que describa de mejor manera el número de reseñas para cada tipo de alojamiento.

Posteriormente se realizarán tablas con todos los coeficientes de determinación y correlación obtenidos para cada tipo de habitación, de cada ciudad.

**Todos los documentos necesarios para realizar el análisis y el desarrollo de la actividad puede descargarse desde el repositorio de mi cuenta personal de Github:*
https://github.com/Luisjagm/Actividad-1

DESARROLLO:

Datos Nulos y Outliers

```
Valores_nulos=df.isnull().sum()
print(Valores_nulos)
df1=df.fillna(method='ffill')
Valores_nulos=df1.isnull().sum()
```

```
print(Valores_nulos)
df1=df1[['room_type','number_of_reviews','host_acceptance_rate','review_scor
es_rating','price','review_scores_cleanliness','availability_365','review_sc
ores_communication']]
Valores_nulos=df1.isnull().sum()
print(Valores_nulos)
```

El bloque de código anterior fue lo que se utilizó para eliminar los datos nulos de las tres tablas de ciudades, realmente el reto de la limpieza de los datos fue el siguiente bloque de código donde al imprimir los tipos de datos que resultaban las columnas se apreciaba que la columna de **price** y de **host_acceptance_rate** estaban consideradas de tipo objeto, esto por el hecho que en los datos se encontraban símbolos de dólar y de porcentaje, por lo tanto, se consideraban strings. Para solucionarlo se escribió el siguiente bloque de código.

```
#Eliminar Signo de Dólar y convertir columna de price a float
df1['price'] = df1['price'].replace({'\$':''}, regex = True)
df1['price'] = df1['price'].replace({',':''}, regex = True)
df1['price'] = df1['price'].astype(str).astype(float)
#Columna de host_acceptance_rate
df1['host_acceptance_rate'] = df1['host_acceptance_rate'].replace({'%':''},
regex = True)
df1['host_acceptance_rate']=df1['host_acceptance_rate'].astype(str).astype(i
nt)
```

Para terminar de preparar los datos se hizo un iterador para que columna por columna obtuviera los valores del primer y tercer cuartil y poder obtener un rango intercuartil, esto para poder detectar outliers y remplazarlos por el valor promedio de la columna correspondiente. Así mismo dividimos en 3 diferentes DataFrames (Private_Room, Shared_Room y Hotel_Room) para poder trabajar por separado con cada uno de los tipos de cuarto. Con esto realizado podemos iniciar nuestro análisis.

```
for columna in df1.drop('room_type', axis=1).columns:
  Q3=df1[columna].quantile(q=0.75)
  Q1=df1[columna].quantile(q=0.25)
  Rq=Q3-Q1
  promedio=df1[columna].mean()
  df1[columna]=df1[columna].mask(df1[columna]<Q1-1.5*Rq, promedio)
  df1[columna]=df1[columna].mask(df1[columna]>Q3+1.5*Rq, promedio)

Private_room=df1[df1['room_type']=='Private room']
Shared_room=df1[df1['room_type']=='Shared room']
Hotel_room=df1[df1['room_type']=='Hotel room']
```

Análisis de Correlación

Para realizar el análisis de correlación utilizamos gráficas estilo “scatterplot” con la librería seaborn referida en el código como “sns” y los resultados obtenidos fueron los siguientes:

```
fig, axes = plt.subplots(2, 3, figsize=(18, 10))
fig.suptitle('PRIVATE ROOM')
sns.scatterplot(ax=axes[0,0],x='host_acceptance_rate',y='number_of_reviews',
color='blue',data=Private_room)
sns.scatterplot(ax=axes[0,1],x='review_scores_rating',y='number_of_reviews',
color='red',data=Private_room)
sns.scatterplot(ax=axes[0,2],x='price',y='number_of_reviews',color='green',d
ata=Private_room)
sns.scatterplot(ax=axes[1,0],x='review_scores_cleanliness',y='number_of_revi
ews',color='blue',data=Private_room)
sns.scatterplot(ax=axes[1,1],x='availability_365',y='number_of_reviews',colo
r='red',data=Private_room)
sns.scatterplot(ax=axes[1,2],x='review_scores_communication',y='number_of_re
views',color='green',data=Private_room)
abs(Private_room.corr())
```

Ciudad de México:

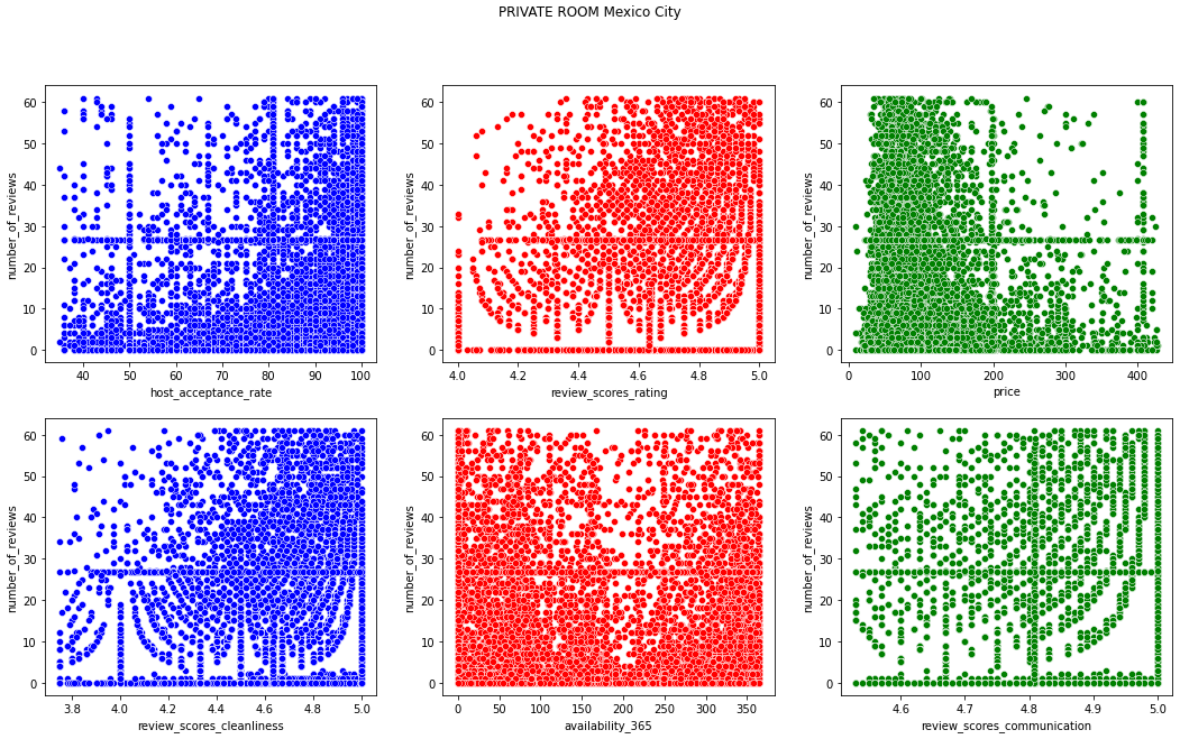

ILUSTRACIÓN 1 SCATTER PLOT PRIVATE ROOM CDMX

	number_of_reviews	host_acceptance_rate	review_scores_rating	price	review_scores_cleanliness	availability_365	review_scores_communication
number_of_reviews	1.000000	0.114019	0.077315	0.103862	0.102857	0.090421	0.141873
host_acceptance_rate	0.114019	1.000000	0.046404	0.081712	0.016612	0.041631	0.042991
review_scores_rating	0.077315	0.046404	1.000000	0.029112	0.603288	0.072713	0.589055
price	0.103862	0.081712	0.029112	1.000000	0.089161	0.123657	0.011276
review_scores_cleanliness	0.102857	0.016612	0.603288	0.089161	1.000000	0.040093	0.461644
availability_365	0.090421	0.041631	0.072713	0.123657	0.040093	1.000000	0.058134
review_scores_communication	0.141873	0.042991	0.589055	0.011276	0.461644	0.058134	1.000000

ILUSTRACIÓN 2 RESULTADOS CORRELACIÓN PRIVATE ROOM CDMX

Chicago:

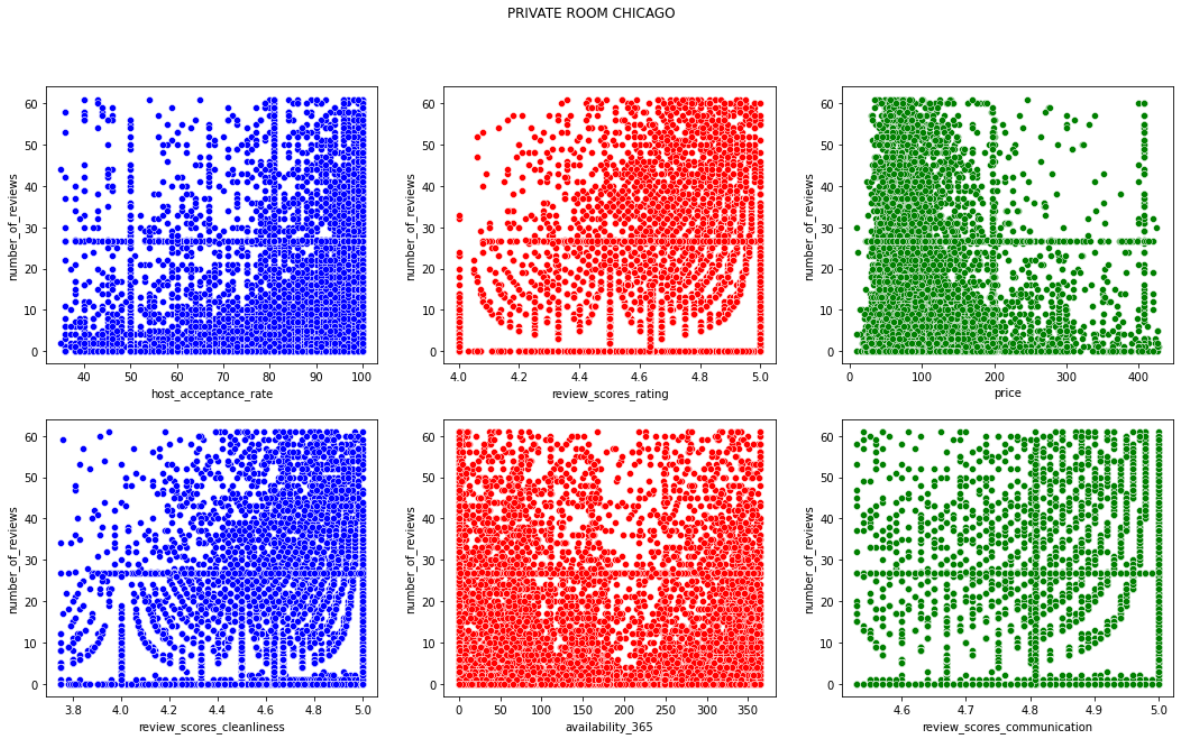

ILUSTRACIÓN 3 SCATTER PLOT PRIVATE ROOM CHICAGO

	Predicciones	number_of_reviews	host_acceptance_rate	review_scores_rating	price	review_scores_cleanliness	availability_365	review_scores_communication
Predicciones	1.000000	1.000000	0.110467	0.033860	0.052642	0.031413	0.118636	0.225381
number_of_reviews	1.000000	1.000000	0.110467	0.033860	0.052642	0.031413	0.118636	0.225381
host_acceptance_rate	0.110467	0.110467	1.000000	0.028032	0.095216	0.017576	0.060859	0.068843
review_scores_rating	0.033860	0.033860	0.028032	1.000000	0.011643	0.560293	0.014890	0.474082
price	0.052642	0.052642	0.095216	0.011643	1.000000	0.050115	0.121293	0.035810
review_scores_cleanliness	0.031413	0.031413	0.017576	0.560293	0.050115	1.000000	0.028627	0.344338
availability_365	0.118636	0.118636	0.060859	0.014890	0.121293	0.028627	1.000000	0.102009
review_scores_communication	0.225381	0.225381	0.068843	0.474082	0.035810	0.344338	0.102009	1.000000

ILUSTRACIÓN 4 RESULTADOS CORRELACIÓN PRIVATE ROOM CHICAGO

Nueva York:

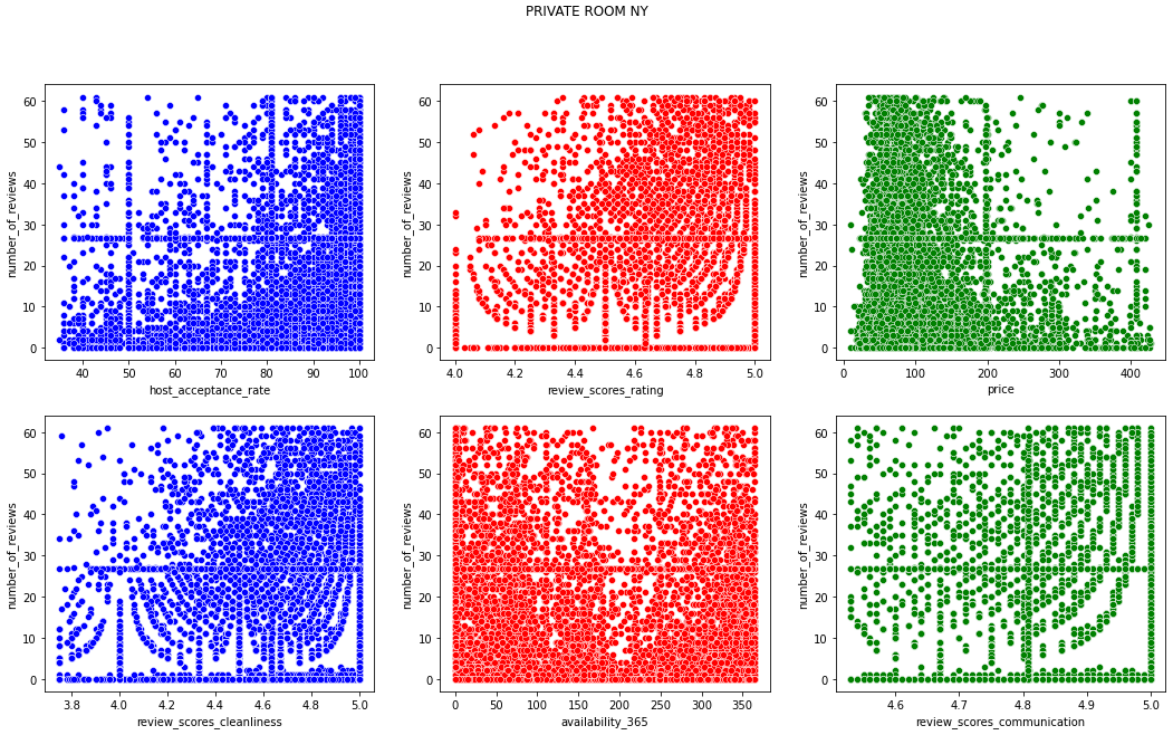

ILUSTRACIÓN 5 SCATTER PLOT PRIVATE ROOM NY

	Predicciones	number_of_reviews	host_acceptance_rate	review_scores_rating	price	review_scores_cleanliness	availability_365	review_scores_communication
Predicciones	1.000000	1.000000	0.110467	0.033860	0.052642	0.031413	0.118636	0.225381
number_of_reviews	1.000000	1.000000	0.110467	0.033860	0.052642	0.031413	0.118636	0.225381
host_acceptance_rate	0.110467	0.110467	1.000000	0.028032	0.095216	0.017576	0.060859	0.068843
review_scores_rating	0.033860	0.033860	0.028032	1.000000	0.011643	0.560293	0.014890	0.474082
price	0.052642	0.052642	0.095216	0.011643	1.000000	0.050115	0.121293	0.035810
review_scores_cleanliness	0.031413	0.031413	0.017576	0.560293	0.050115	1.000000	0.028627	0.344338
availability_365	0.118636	0.118636	0.060859	0.014890	0.121293	0.028627	1.000000	0.102009
review_scores_communication	0.225381	0.225381	0.068843	0.474082	0.035810	0.344338	0.102009	1.000000

ILUSTRACIÓN 6 RESULTADOS CORRELACIÓN PRIVATE ROOM NY

Como se puede observar las 3 gráficas son prácticamente idénticas, pero si nos fijamos en las tablas de correlación podemos confirmar que son totalmente diferentes. La variable de respuesta es "number of reviews" y comparando los valores de correlación, en cada uno de los análisis coincide que la mayor correlación existente es con la característica: "review_scores_communication" pero en los 3 casos podemos decir que se presenta una correlación positiva débil por lo tanto podemos esperar que nuestro modelo matemático no sea muy preciso.

Utilizamos como ejemplo al DataFrame Private_Room para demostrar el procedimiento del análisis los demás análisis de Shared_Room y Hotel_Room se encuentra en el cuaderno de notas de Jupyter en el repositorio de Github.

Aunque cabe destacar que de los 3 DataFrames analizados la única correlación que se pudo considerar como "positiva media" (+0.50) fue en los datos de Nueva York y en Hotel_Room y con la misma variable "review_scores_communication", este fue el mejor resultado obtenido, pero sigue siendo muy bajo el resultado.

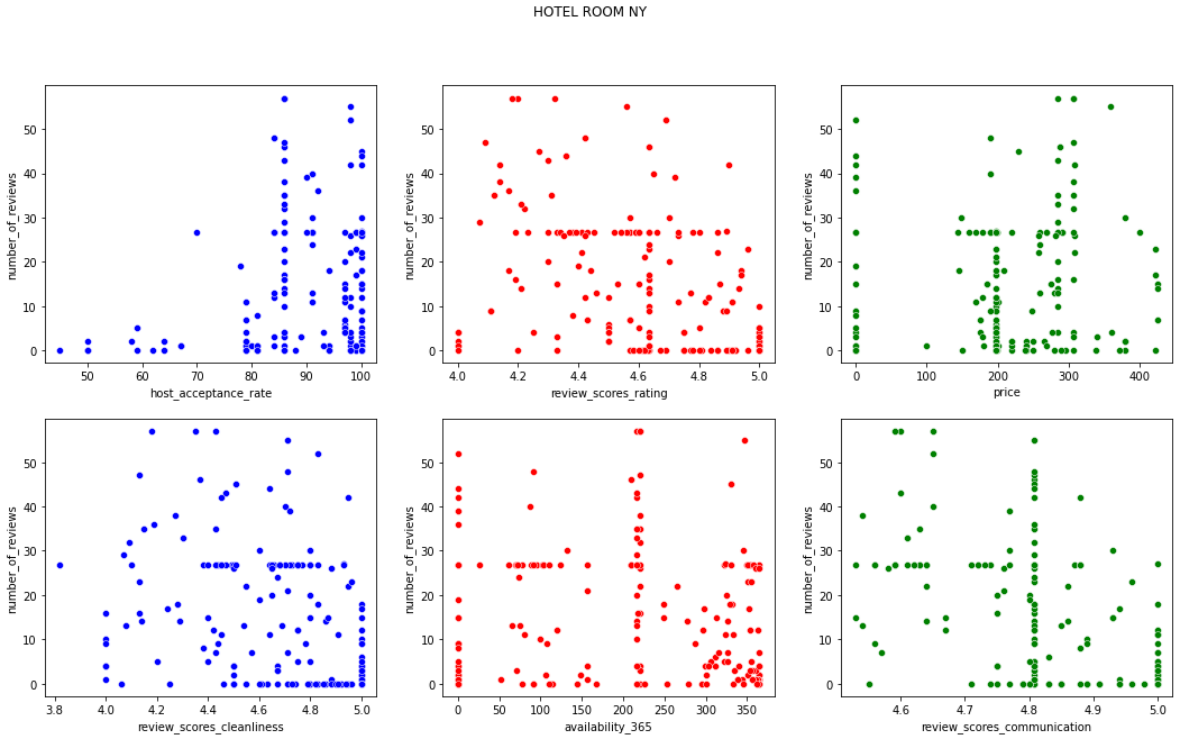

ILUSTRACIÓN 7 SCATTER PLOT HOTEL ROOM NY

	Predicciones	number_of_reviews	host_acceptance_rate	review_scores_rating	price	review_scores_cleanliness	availability_365	review_scores_communication
Predicciones	1.000000	1.000000	0.086405	0.460889	0.145465	0.410524	0.019850	0.562974
number_of_reviews	1.000000	1.000000	0.086405	0.460889	0.145465	0.410524	0.019850	0.562974
host_acceptance_rate	0.086405	0.086405	1.000000	0.003157	0.046559	0.044406	0.204846	0.021424
review_scores_rating	0.460889	0.460889	0.003157	1.000000	0.119131	0.646870	0.071783	0.566936
price	0.145465	0.145465	0.046559	0.119131	1.000000	0.068308	0.447615	0.047211
review_scores_cleanliness	0.410524	0.410524	0.044406	0.646870	0.068308	1.000000	0.088489	0.479995
availability_365	0.019850	0.019850	0.204846	0.071783	0.447615	0.088489	1.000000	0.039525
review_scores_communication	0.562974	0.562974	0.021424	0.566936	0.047211	0.479995	0.039525	1.000000

ILUSTRACIÓN 8 RESULTADOS CORRELACIÓN HOTEL ROOM NY

GENERACIÓN DEL MODELO MATEMÁTICO

```
#Declaramos las variables dependientes e independientes para la regresión
lineal
from sklearn.linear_model import LinearRegression
modelSR=LinearRegression()
SRVars_Indep= Shared_room[['review_scores_communication']]
SRVar_Dep= Shared_room['number_of_reviews']
modelSR.fit(X=SRVars_Indep,y=SRVar_Dep)
y_pred=modelSR.predict(X=Shared_room[['number_of_reviews']])
```

En el bloque de código anterior se puede observar como fue la creación del modelo matemático para cada una de los tipos de cuarto y su ciudad correspondiente. Así mismo se realizó la predicción de la variable de respuesta y de agregó la columna de predicciones al DataFrame del tipo de cuarto correspondiente.

Tablas de coeficientes de determinación y correlación

Para poder determinar los coeficientes de determinación y correlación para cada una de las ciudades se utilizó el siguiente bloque de código:

```
Coef_Deter_Hotel=round(modelHR.score(Vars_Indep,Var_Dep),5)
Coef_Deter_Shared=round(modelSR.score(Vars_Indep,Var_Dep),5)
Coef_Deter_Private=round(modelPR.score(Vars_Indep,Var_Dep),5)
Coef_Corr_Hotel=round(np.sqrt(abs(Coef_Deter_Hotel)),5)
Coef_Corr_Shared=round(np.sqrt(abs(Coef_Deter_Shared)),5)
Coef_Corr_Private=round(np.sqrt(Coef_Deter_Private),5)
Coefs = ["Coef_Deter", "Coef_Corr"]
titulo = ["Private_Room", "Shared_Room", "Hotel_Room"]
data = [[Coef_Deter_Private, Coef_Deter_Shared, Coef_Deter_Hotel],
[Coef_Corr_Private, Coef_Corr_Shared, Coef_Corr_Hotel]]
format_row = "{:>12}" * (len(titulo) + 1)
print(format_row.format("", *titulo))
for team, row in zip(Coefs, data):
    print(format_row.format(team, *row))
```

Esto aparte imprime las 3 tablas de las ciudades, por fines de practicidad se mostrarán los resultados en las siguientes tablas:

Ciudad de México

	Private_Room	Shared_Room	Hotel_Room
Coeficiente de Determinación	0.02013	-0.14644	-0.1464
Coeficiente de Correlación	0.14188	0.38267	0.38267

Nueva York

	Private_Room	Shared_Room	Hotel_Room
Coeficiente de Determinación	0.0508	0.0342	0.0342
Coeficiente de Correlación	0.22539	0.18493	0.18493

Chicago

	Private_Room	Shared_Room	Hotel_Room
Coeficiente de Determinación	0.00572	-0.08311	-0.08311
Coeficiente de Correlación	0.07563	0.28829	0.28829