Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Querétaro

Actividad 1: Reporte Comparativo de Resultados

Analítica de Datos y Herramientas de Inteligencia Artificial II
Grupo 502

Keyla Patricia Islas Garrido A01730349

Profesor: Alfredo García Suárez

Tabla de Contenidos

Airbnb (*Airbed and breakfast*) es una compañía estadounidense que innovó en la industria del alojamiento al poner a disposición una plataforma digital dedicada a la oferta de arrendamiento de alojamientos en diversas ciudades y países alrededor del mundo ofreciendo nuevas y diversas opciones a anfitriones y huéspedes. Esta compañía afirma ser parte de la "economía colaborativa" y "sacudir" la industria hotelera. Sin embargo, los datos muestran que la mayoría de los listados de Airbnb en la mayoría de las ciudades son casas completas, muchas de las cuales se alquilan durante todo el año, lo que altera en realidad a las viviendas y las comunidades.

En estadística el concepto de *Regresión Lineal Simple* se refiere a una técnica de análisis de datos que consiste en predecir una variable objetivo dependiente de una variable explicativa con la que se relaciona a través de una ecuación lineal. Es un modelo de fácil comprensión y sencillo por lo que se ha vuelto un modelo de uso común en el estudio de relaciones causa-efecto entre dos variables.

Objetivos

Para propósitos de evaluación de aprendizaje se plantea usar esta base de datos sobre tres ciudades: *D.F., México, Atenas, Grecia, Ámsterdam, Países Bajos* a fin de abismar en el tema de *Regresión Lineal Simple* y sus componentes principales *modelo de regresión, correlación, coeficiente de Pearson, coeficiente de Determinación,* entre otros.

El objetivo principal es crear un modelo de regresión lineal simple que sea útil para la predicción de *Número de Reviews* para cada tipo de habitación ofrecida (*Entire home/apt, Share Room, Private Room* y *Hotel Room*).

Metodología

Extracción de Datos

Desde el sitio online *Inside Airbnb* se ponen a disposición pública datos depurados provenientes de información disponible públicamente en el sitio online oficial de *Airbnb*. Dentro de los repositorios de datos pertenecientes a docenas de ciudades alrededor del mundo se encuentra el repositorio *listings* que nos brinda información detallada sobre los lugares ofertados en su plataforma digital. Se extrajeron las bases de datos *listings* correspondientes a las ciudades de *D.F., México, Atenas, Grecia* y *Ámsterdam, Países Bajos.*

Limpieza de Datos: Valores Nulos y Atípicos

La limpieza de datos cosiste en:

1) La selección de las variables de interés.
   - *Host Acceptance Rate* (Tasa de Aceptación de Huéspedes)
   - *Price* (Precio)
   - *Availability 365* (Días de Disponibilidad Futuros (365 días en el futuro))
   - *Review Scores Rating* (Calificación en Reseñas: Clasificación)
   - *Review Scores Cleanliness* (Calificación en Reseñas: Higiene)
   - *Review Scores Communication* (Calificación en Reseñas: Comunicación)
2) Omisión de caracteres no admisibles en variables numéricas (,,$,%) y su conversión.

```
for i in dff.columns[1:]:
  if (dff[i].dtype == 'object'):
    for x,y in zip(dff[i], range(len(dff))):
      if (x!=NaN):
        z = str(x)
        z = z.replace('%','')
        z = z.replace(',','')
        z = z.replace('$','')
        dff[i][y] = z
    dff[i] = pd.to_numeric(dff[i], errors = 'coerce')
```

3) Tratamiento de valores nulos
   Debido a que las variables de interés son numéricas, los valores nulos fueron reemplazados por la media.

```
for i in dff.columns[1:]:
  if (dff[i].dtype == 'int64'):
    dff[i].fillna(round(dff[i].mean()), inplace = True)
  if (dff[i].dtype == 'float64'):
    dff[i].fillna(round(dff[i].mean(),2), inplace = True)
```

4) Tratamiento de valores atípicos
   Se usa el método de la Campana de Gauss y la distribución normal para detectar valores atípicos y después reemplazarlos por la media.

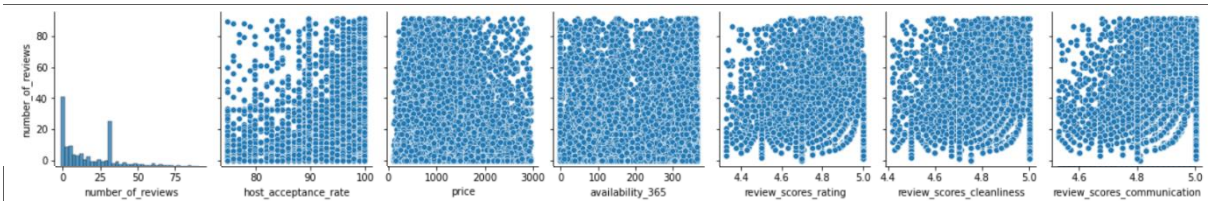
```
p25 = dff.quantile(0.25)
p75 = dff.quantile(0.75)
iqr = p75-p25
ls = p75 + 1.5*iqr
li = p25 - 1.5*iqr
dff1 = dff.loc[:, dff.columns != 'room_type'][(dff<=ls)&(
dff>=li)]
dff1.fillna(round(dff.mean(),1), inplace = True)
dff1 = pd.concat([dff['room_type'], dff1], axis=1)
```

Diagramas de Dispersión y Mapa de Calor Correlación

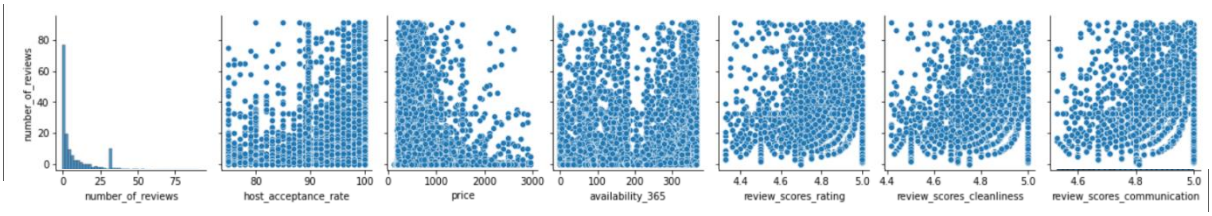
Para un mejor manejo del problema, las bases de datos de cada ciudad fueron divididas en 4 dependiendo del tipo de habitación ofrecida dando un total de 12 datasets.

Se muestran los diagramas de dispersión obtenidos de esta división solo para los tipos de habitación en D.F., México.

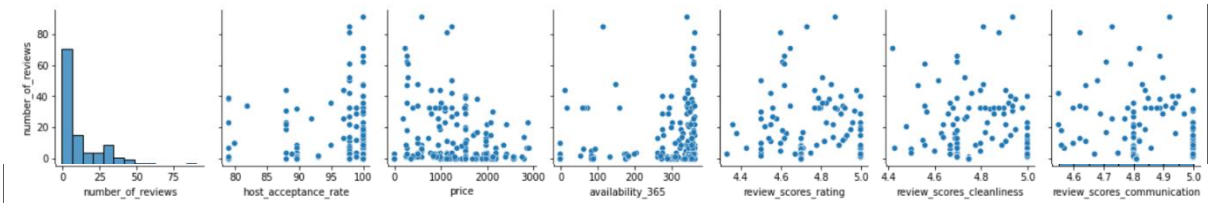
- D.F., México – Entire home/apt



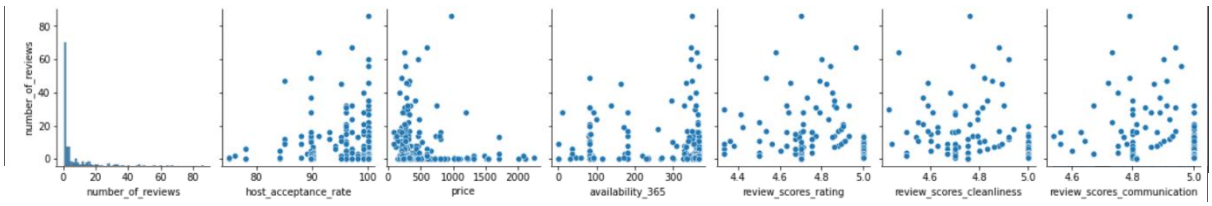
- D.F., México – Private Room



- D.F., México – Shared Room



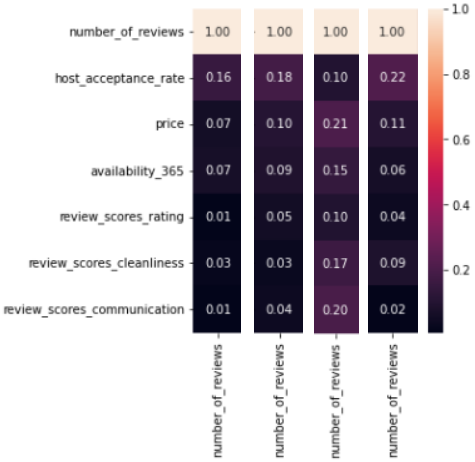
- D.F., México – Hotel Room

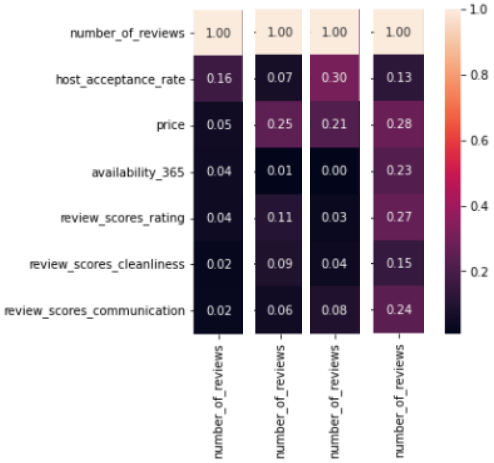


En base a estos diagramas, se expresa que las variables de interés no tienen una correlación significativa entre sí, se espera que la correlación máxima es de R=0.2 para variables que muestran un pequeño comportamiento ascendente como *host_acceptance_rate* y la mínima de R = 0 para variables como *review_scores* que se muestran muy dispersas para el caso de las habitaciones en D.F.

Esto se puede comprobar con un mapa de calor de los coeficientes de correlación que se muestra a continuación en orden *Entire home/apt*, *Private Room*, *Shared Room* y *Hotel Room*:

Se concluye que las variables con mejor correlación para cada tipo de habitación son *host_acceptance_rate*, *host_acceptance_rate*, *Price* y *host_acceptance_rate* respectivamente.

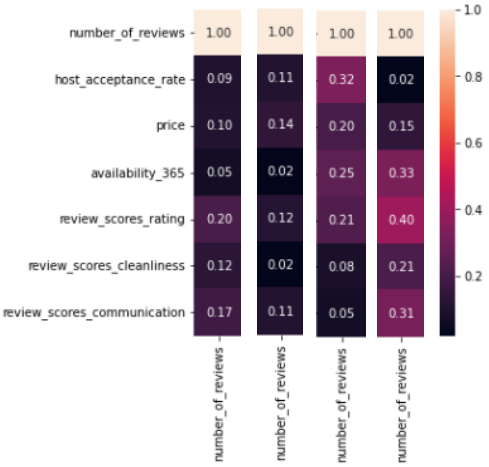




En caso de Atena, Grecia el mapa de calor para la correlación de sus variables por tipo de habitación en el mismo orden es el siguiente:

Para *Entire home/apt* la variable con mejor correlación es *host_acceptance_rate*; para *Private Room*, *Price*; para *Shared Room*, *host_acceptance_rate* con un R=0.3 que excede a lo visto en los datos de D.F.; *Hotel Room* en Atenas es el tipo de habitación que tiene mejores variables hasta el momento, pero aun con una correlación poco significativa con la mejor siendo *Price*.

En el caso de Ámsterdam, Países Bajos el mapa de calor de correlaciones se muestra así con el tipo de habitación en el mismo orden *Entire home/apt*, *Private room*, *Shared room*, *Hotel room* y las variables que muestran mejor correlación son *review_scores_rating* con un R=0.2, *price* con un R= 0.14, *host_acceptance_rate* con R= 0.32 que vuelve a superar los registros anteriores y *review_scores_rating* con R=0.40 que es la variable que tiene mejor correlación con *number of reviews* de los 12 datasets usando *Hotel Room en Ámsterdam*



Modelo de Regresión Lineal Simple

Para crear el modelo de Regresión Lineal Simple se hace uso de la librería *sklearn* que nos permite usar la función *LinearRegression* para encontrar una intersección y una pendiente lineal que relacione la variable dependiente con la independiente.

```
1. from sklearn.linear_model import LinearRegression
2. model= LinearRegression()
3. model.fit(X=Var_Independiente, y=Var_Dependiente)
4. print('El mejor modelo para el dataset es: y =
   ', model.__dict__['coef_'][0],'*Var_Independiente+
   ',model.__dict__['intercept_'])
```

A continuación, se muestran los modelos hechos para los tipos de habitación en D.F., Atenas y Ámsterdam considerando las variables con mejor correlación determinadas en el punto anterior con el propósito de predecir el número de reseñas hechas al lugar de alojamiento.

- D.F., México

```
El mejor modelo para el dataset  Entire home/apt  es: y
=  0.6348975678360113 *host_acceptance_rate  -39.31343139108138

El mejor modelo para el dataset  Private room  es: y
=  0.573027828056219 *host_acceptance_rate  -40.914670511309666

El mejor modelo para el dataset  Shared room  es: y
=  0.4682254177860051 *host_acceptance_rate  -38.31766709398862

El mejor modelo para el dataset  Hotel room  es: y =  -
0.0053727625635090325 *price +  20.621597417581462
```

- Atenas, Grecia

```
El mejor modelo para el dataset  Entire home/apt  es: y
=  1.2263425275921414 *host_acceptance_rate  -92.95923926028586

El mejor modelo para el dataset  Hotel room  es: y
=  1.6449025719864836 *host_acceptance_rate  -147.86096276087665

El mejor modelo para el dataset  Private room  es: y =  -
0.13088483941955886 *price + 21.861597908745132

El mejor modelo para el dataset  Shared room  es: y =  -
0.25967148246986294 *price + 15.797983524111693
```

- Ámsterdam, Países Bajos

```
El mejor modelo para el dataset  Entire home/apt  es: y =  -
31.57831362541099 *review_scores_rating + 172.1645958901621

El mejor modelo para el dataset  Shared room  es: y =  -
72.62023927662278 *review_scores_rating + 380.58247277165174

El mejor modelo para el dataset  Private room  es: y =  -
0.04574847907224517 *price +  46.05922369804776

El mejor modelo para el dataset  Hotel room  es: y
=  1.1161793850048911 *host_acceptance_rate  -71.7480499420837
```

Coeficientes de Correlación y Determinación

El coeficiente de correlación (R) es aquel que nos explica que tan fuerte es la relación lineal entre dos variables, puede ser positivo o negativo, mientras ente más cerca de -1 o 1 representa una relación fuerte por otro lado, mientras más cerca esta de 0 menos relación hay entre las variables. Este coeficiente fungió como base para escoger las variables que harían mejor un modelo de regresión lineal simple.

Su valor al cuadrado es el coeficiente de determinación (R^2) representa cuanta varianza explica el modelo o, en otras palabras, indica que tan perfecto es el ajuste del modelo y la capacidad de explicación de la recta. Su valor va de 0 a 1, entre más cercano a 1, mejor ajuste tiene el modelo y mientras más cercano a 0, la recta tendrá baja capacidad de explicación/predicción.

El calculo del coeficiente de determinación forma parte tambien de las funciones dadas por la librería *sklearn* y la función *LinearRegression*. Sin embargo, para el calculo del coeficiente de correlación se debe sacar la raíz cuadrada por lo que no se sabrá solo con esto si la relación es directa o inversamente proporcional como podría verse con la función *corr()* usada para crear los mapas de calor de la correlación entre variables.

```
1. #Coeficiente de Determinación
2. model.score(var_Independiente, Var_Dependiente)
3. #Coeficiente de Correlación
4. np.sqrt(model.score(var_Independiente, Var_Dependiente))
```

A continuación, se muestran las tablas de coeficientes resultantes de todo el proceso.

- D.F. México

Tipo de habitación	Entire home/apt	Private room	Shared room	Hotel room
Variables con mayor correlación	Host Acceptance Rate	Host Acceptance Rate	Host Acceptance Rate	Price
Coeficiente de Correlación (R)	0.158501	0.180329	0.215574	0.213049
Coeficiente de Determinación (R^2)	0.025123	0.032518	0.046472	0.04539

- Atenas, Grecia

Tipo de habitación	Entire home/apt	Hotel room	Private room	Shared room
Variables con mayor correlación	Host Acceptance Rate	Host Acceptance Rate	Price	Price
Coeficiente de Correlación (R)	0.162226	0.303546	0.252706	0.280162
Coeficiente de Determinación (R^2)	0.026317	0.09214	0.06386	0.078491

- Ámsterdam, Países Bajos

Tipo de habitación	Entire home/apt	Shared room	Private room	Hotel room
Variables con mayor correlación	Review Scores Rating	Review Scores Rating	Price	Host Acceptance Rate
Coeficiente de Correlación (R)	0.19784	0.404931	0.143984	0.323867
Coeficiente de Determinación (R^2)	0.039141	0.163969	0.020731	0.10489

Hallazgos

El coeficiente de correlación más alto encontrado fue de R= 0.4, un tipo de correlación débil, siendo el mismo caso para los demás coeficientes representativos por lo que a pesar de ser los mejores modelos que pudieron crearse, no son buenos modelos y esto se confirma con los coeficientes de determinación debajo de $R^2 = 0.2$.

En general, la correlación es mejor dependiendo de la ciudad; Ámsterdam tiene mejores coeficientes de correlación que Atenas y los de Atenas son mejores que los de D.F.

Las variables representativas que se presentan en las tres ciudades son la tasa de aceptación de huéspedes y el precio; de estas dos la tasa de aceptación de huéspedes es la más recurrente ya que fue la mejor opción en 6 de 12 data sets mientras que la variable de precio fue la mejor opción en 4 de 12 data sets. En general, para que haya más reseñas la tasa de aceptación de huéspedes debe ser alta y el precio debe ser bajo. Es decir, un lugar con más reseñas tuvo más huéspedes alojados en ella a un precio bajo. Esto a su vez, representa causalidad.

De acuerdo con los resultados, la cantidad de reseñas para las casas o departamentos completos depende de la buena tasa de aceptación de huéspedes para las ciudades de D.F. y Atenas mientras que para Ámsterdam lo importante es la clasificación dada en las reseñas.

En caso de las habitaciones privadas, se tendrán más reseñas si el precio fue económico para las ciudades de Atenas y Ámsterdam mientras que para D.F. lo importante es la tasa de aceptación de huéspedes.

Para las habitaciones compartidas cada ciudad tiene una variable representativa diferente. Para D.F. habrá más reseñas dependiendo de la tasa de aceptación de huéspedes, en Atenas depende del precio y en Ámsterdam depende de la clasificación en reseñas.

Por último, para las habitaciones de hotel en Atenas y Ámsterdam la cantidad de reseñas dependerá de la tasa de aceptación de huéspedes mientras que para D.F. depende de un precio accesible.