Regina Abedrop Payró

A01702211

Actividad 1

En el siguiente análisis se muestra la evaluación de distintos modelos de regresión lineal y múltiple. Para los modelos de regresión lineal se utilizó la base de datos Listing.csv.gz de México obtenido de <http://insideairbnb.com/get-the-data/> y para los modelos de regresión lineal múltiples se utilizaron las bases de datos de Listing.csv.gz de Atenas y Listing.csv.gz de Bangkok obtenidos del mismo enlace mencionado anteriormente.

**REGRESION LINEAL**

**Para iniciar el análisis lo primero que realice fue** cargar los archivos con las bases de datos de México a Python en Google Colab y realice un preprocesamiento de los datos en donde deje separe la base de datos original en 3 nuevas DataFrames (uno por cada tipo de habitación: Entire House, Private Room y Hotel Room) dejando únicamente las columnas numéricas que considere relevantes para realizar el modelo. Posteriormente verifique que no hubiera datos nulos o outliers reemplazando valores nulos y utilizando el método de cuartiles para tratar con los datos atípicos de las bases de datos.

Para iniciar el análisis es importante mencionar que nuestra variable dependiente para todos los análisis es el “Number of Reviews” y las variables independientes son “host\_acceptance\_rate”, “price”, “availability\_365”, “review\_scores\_rating”, “review\_scores\_cleanliness” y “review\_scores\_communication”.

**Análisis “Entire House”**

Al realizar el análisis únicamente de Entire House podemos observar que hay una gran dispersión de puntos en las graficas que se muestran a continuación, esto quiere decir que no hay una correlación fuerte entre las independientes con nuestra dependiente.

Chart, scatter chart

Description automatically generated

En el siguiente mapa de color podemos observar que hay una relación muy débil entre todas las variables, al fijarnos en las correlaciones que las variables independientes tienen con nuestra variable dependiente podemos observar que son de colores muy claro lo que representa que la correlación es muy baja.

Chart, histogram

Description automatically generated

La siguientes graficas muestran los diagramas de dispersión y en puntos negros muestra el modelo que creamos, podemos observar que la mayoría de los puntos no se ajustan al modelo que se creó.

Chart, scatter chart

Description automatically generated

**Análisis “Private Room”**

Al realizar el análisis únicamente de Private Room podemos observar que hay una gran dispersión de puntos en las gráficas que se muestran a continuación, esto quiere decir que no hay una correlación fuerte entre las independientes con nuestra dependiente.

Chart, scatter chart

Description automatically generated

En el siguiente mapa de calor podemos observar que hay una relación muy débil entre todas las variables, al fijarnos en las correlaciones que las variables independientes tienen con nuestra variable dependiente podemos observar que son de colores muy claro lo que representa que la correlación es muy baja.

Chart, histogram

Description automatically generated

La siguientes graficas muestran los diagramas de dispersión y en puntos negros muestra el modelo que creamos, podemos observar que la mayoría de los puntos no se ajustan al modelo que se creó.

Chart, scatter chart

Description automatically generated

**Análisis “Hotel room”**

Al realizar el análisis únicamente de Hotel Room podemos observar que hay una gran dispersión de puntos en las gráficas que se muestran a continuación, esto quiere decir que no hay una correlación fuerte entre las independientes con nuestra dependiente.

Chart, scatter chart

Description automatically generated

En el siguiente mapa de calor podemos observar que hay una relación muy débil entre todas las variables, al fijarnos en las correlaciones que las variables independientes tienen con nuestra variable dependiente podemos observar que son de colores muy claro lo que representa que la correlación es muy baja.

Chart, histogram

Description automatically generated

La siguientes graficas muestran los diagramas de dispersión y en puntos negros muestra el modelo que creamos, podemos observar que la mayoría de los puntos no se ajustan al modelo que se creó.

Chart, scatter chart

Description automatically generated

Tablas comparativas:

|  |  |  |  |
| --- | --- | --- | --- |
| **Coeficientes de determinacion** | | | |
|  | **Entire Home** | **Private Room** | **Hotel Room** |
| **Modelo 1** | 0.01574353 | 0.016467286 | 0.02541239 |
| **Modelo 2** | 0.00392207 | 0.004763529 | 0.00556822 |
| **Modelo 3** | 0.00413475 | 0.009095158 | 0.00962094 |
| **Modelo 4** | 0.00266332 | 0.005204271 | 0.00729684 |
| **Modelo 5** | 0.00630443 | 0.00708563 | 6.6661E-05 |
| **Modelo 6** | 0.004913 | 0.002753366 | 0.00686601 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coeficientes de correlacion con la variable Number of reviews** | | | |
|  | **Entire Home** | **Private Room** | **Hotel Room** |
| **host\_acceptance\_rate** | 0.125473 | 0.128325 | 0.159413 |
| **price** | -0.062626 | -0.069018 | -0.074621 |
| **availability\_365** | -0.064302 | -0.095369 | 0.098086 |
| **review\_scores\_rating** | 0.051607 | 0.072141 | 0.085422 |
| **review\_scores\_cleanliness** | 0.0794 | 0.084176 | -0.008165 |
| **review\_scores\_communication** | 0.070093 | 0.052473 | -0.082861 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Coeficientes de correlacion de las predicciones con la variable Number of reviews** | | | |
|  | **Entire Home** | **Private Room** | **Hotel Room** |
| **Predicciones 1** | 0.125473 | 0.128325 | 0.159413 |
| **Predicciones 2** | 0.062626 | 0.069018 | 0.074621 |
| **Predicciones 3** | 0.064302 | 0.095369 | 0.098086 |
| **Predicciones 4** | 0.051607 | 0.072141 | 0.085422 |
| **Predicciones 5** | 0.0794 | 0.084176 | 0.008165 |
| **Predicciones 6** | 0.070093 | 0.052473 | 0.082861 |

**REGRESION LINEAL MULTIPLE**

Al igual que en la sección de regresión lineal, en esta sección lo primero que realice fue cargar los archivos con las bases de datos de Atenas y de Bangkok a Python en Google Colab y realice un preprocesamiento de los datos en donde separe las bases de datos para trabajar con un solo tipo de cuarto, en este caso únicamente trabaje con el cuarto “Entire Home” y deje únicamente las columnas numéricas que considere relevantes para realizar el modelo, posteriormente verifique que no hubiera datos nulos o outliers reemplazando valores nulos y utilizando el método de cuartiles para tratar con los outliers de las bases de datos.

Para realizar el análisis de regresión lineal múltiple utilice la variable de salida (y = “Number of reviews”) y las variables predictoras (X= 'minimum\_nights', 'number\_of\_reviews', 'price', 'availability\_365', 'calculated\_host\_listings\_count', 'room\_type').

El coeficiente de determinación obtenido en el análisis de la ciudad de Atenas fue de 0.0073 y el de la ciudad de Bangkok fue de 0.0156. Ambos son números muy bajos lo que nos indica que no es un buen modelo.

En conclusión, en todos los análisis obtuvimos resultados muy bajos con los modelos que creamos para predecir la variable y (“Number of reviews”). Al utilizar regresión lineal simple y regresión lineal multiple no logre obtener un modelo que se ajustara a los datos y que lograra tener buenas predicciones, esto se debe a que las correlaciones entre variables son muy bajas es decir que es muy complicado realizar predicciones ya que no se relacionan de manera significativa entre ellas.

El peor modelo de las 3 ciudades fue el de Atenas ya que el coeficiente de determinación fue el más bajo de todos los modelos con un valor de 0.0073.