Adrián Urabayen Arriaga

A01733911

23-10-2022

Reporte Actividad 1

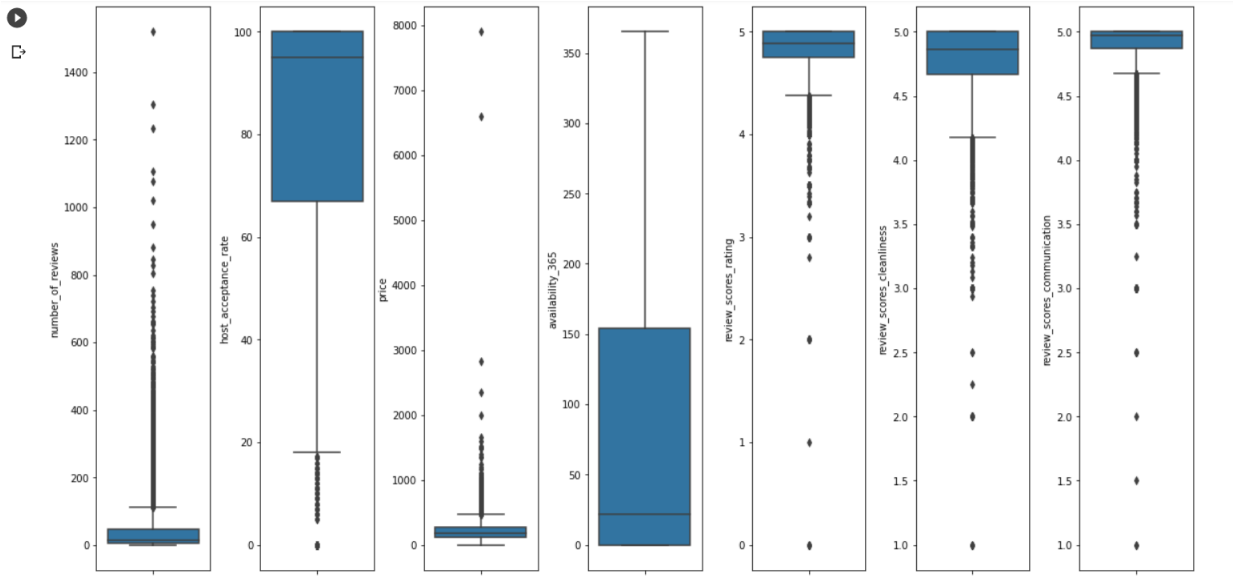
Regresión lineal simple con el df original:

Limpieza y preparación de datos:

```
#Limpieza de filas repetidas
df.drop_duplicates(keep='first', inplace=True)
df.shape
```

```
df =df.fillna(method="bfill")
df =df.fillna(method="ffill")
df.isnull().sum()
```

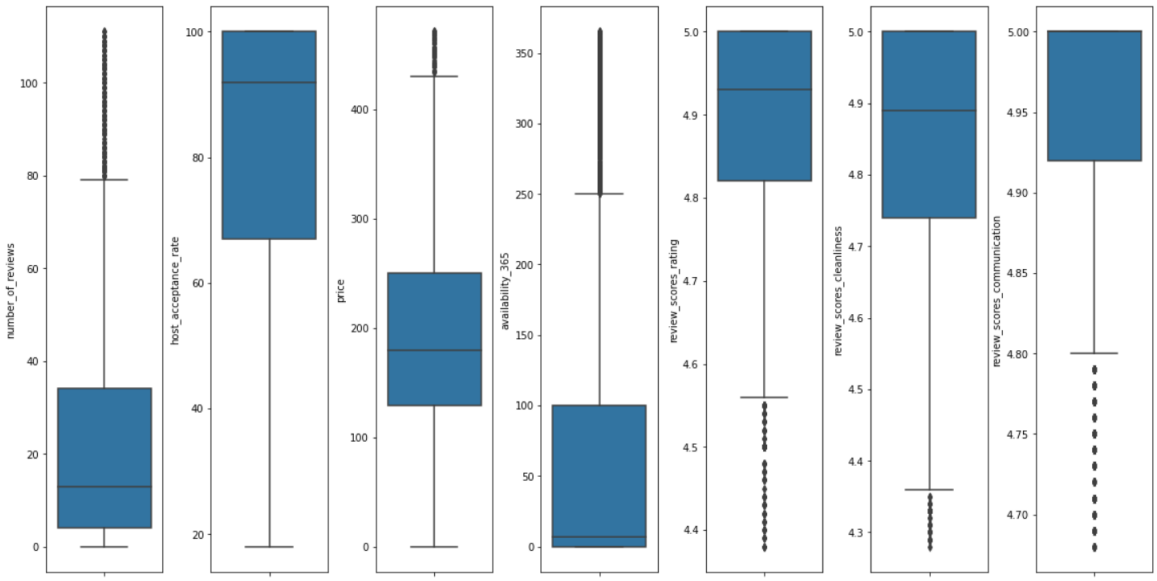
Para la limpieza y preparación de datos, se seleccionaron solo las variables relevantes del dataset y se introdujeron en un dataframe. A continuación, con las funciones mostradas en las imágenes superiores, se eliminaron las filas duplicadas y se sustituyeron los datos atípicos mediante el método backfill y forwardfill.



```
Limite superior permitido number_of_reviews             111.500
host_acceptance_rate          149.500
price                         472.500
availability_365              385.000
review_scores_rating            5.375
review_scores_cleanliness       5.495
review_scores_communication     5.195
dtype: float64
Limite inferior permitido number_of_reviews             -60.500
host_acceptance_rate           17.500
price                         -83.500
availability_365             -231.000
review_scores_rating            4.375
review_scores_cleanliness       4.175
review_scores_communication     4.675
dtype: float64
```

```
df=df[df['number_of_reviews']<111.500]
df=df[df['host_acceptance_rate']>17.500]
df=df[df['price']<472.500]
df=df[df['review_scores_rating']>4.375]
df=df[df['review_scores_cleanliness']>4.275]
df=df[df['review_scores_communication']>4.675]
df.head()
```

Como podemos observar en el primer Boxplot, teníamos una gran cantidad de datos atípicos, los cuales entorpecerían nuestro análisis. Por lo tanto, se decidió eliminarlos. Para ello se calcularon límites superiores e inferiores como se muestran en la imagen anterior y se realizaron filtros entorno a esos límites obtenidos.



De esta manera se volvieron a realizar los boxplots, los cuales dieron unos resultados mejores como se puede observar.

Una vez hecha esta limpieza de datos, se dividió el dataframe por tipos de habitación y para cada uno de estos, se procedió a realizar modelos de regresión lineal simple teniendo en cuenta la variable con mayor correlación obtenida con la variable dependiente 'number_of_reviews' y tomándola como nuestra variable independiente. Finalmente se registraron los resultados de los coeficientes de determinación y correlación en una tabla:

	Tipo de habitacion	Coef_det	Coef_Correl
0	Entire room/apt	0.036467	0.190962
1	Private room	0.047640	0.218266
2	Shared room	0.103115	0.321115
3	Hotel room	0.103981	0.322461

Finalmente, se realizó el mismo proceso mencionado para dos datasets diferentes de otras dos ciudades, en este caso fueron Barcelona y Austin. Sin embargo, en lugar de realizar una regresión lineal simple, se hizo una regresión lineal múltiple, teniendo en cuenta todas las variables de relevancia. Y estos fueron los resultados obtenidos:

Regresión lineal múltiple con df de Barcelona:

	Tipo de habitacion	Coef_det	Coef_Correl
0	Entire room/apt	0.099410	0.315293
1	Private room	0.033537	0.183132
2	Shared room	0.039611	0.199024
3	Hotel room	0.199522	0.446679

Regresión lineal múltiple con df de Austin:

	Tipo de habitacion	Coef_det	Coef_Correl
0	Entire room/apt	0.056852	0.238437
1	Private room	0.068696	0.262099
2	Shared room	0.249064	0.499063
3	Hotel room	0.615317	0.784422

Una vez realizados los 3 diferentes análisis, podemos llegar a la conclusión general de que los modelos de regresión múltiple fueron mejores que el de regresión lineal simple, centrándonos en

los resultados. Sin embargo, debido a la diferencia de los datos de cada dataset no se puede hacer una comparación óptima y realizar un reporte comparativo a profundidad.