José Luis Spíndola Gómez

Actividad 1: Regresión Lineal Simple y Múltiple

Para empezar trabajé con los datos nulos y outliers para que la base estuviera completa y limpia, lista para el análisis.

4)

Host acceptance rate vs host response rate

Entire home	0.53
Private room	0.54
Hotel	0.08
Shared room	0.43

Host acceptance rate vs price

Entire home	0.05
Private room	0.06
Hotel	0.04
Shared room	0.10

Host acceptance rate vs number of reviews

Entire home	0.19
Private room	0.12
Hotel	0.20
Shared room	0.15

Reviews score rating vs calculated host listings count

Entire home	0.04
Private room	0.02
Hotel	0.5

Shared room	0.16

Availability 365 vs number of reviews

Entire home	0.03
Private room	0.06
Hotel	0.12
Shared room	0.27

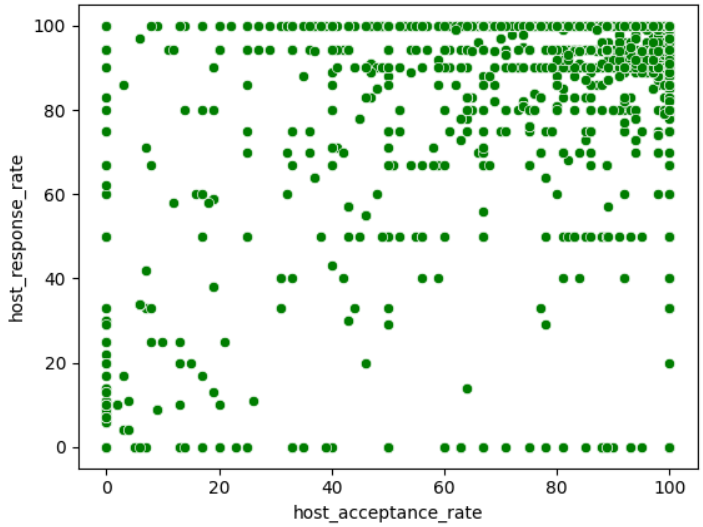
Reviews per month vs review scores communication

Entire home	0.96
Private room	0.97
Hotel	0.99
Shared room	0.97

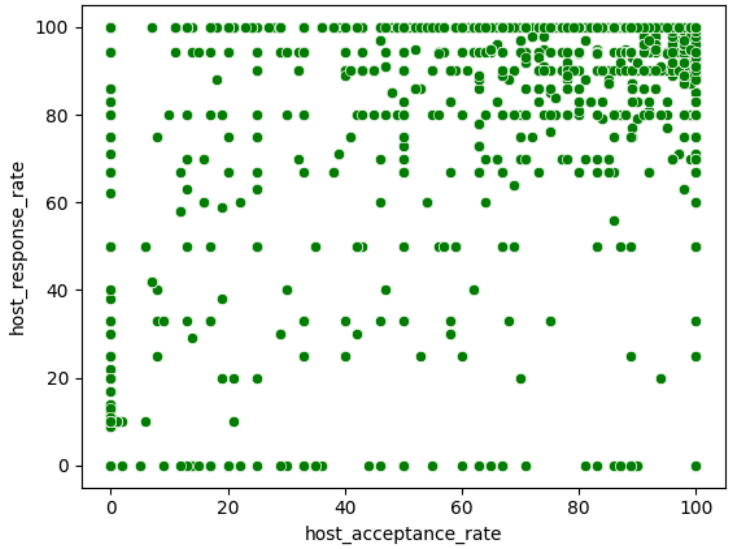
Decidí enfocar el análisis en availability 365 vs number of reviews ya que existe una gran variación que va desde 0.03 para entire home hasta 0.27 para shared room.

En estos se puede apreciar que la tasa de aceptación se concentra en los porcentajes más elevados, aunque en entire home y private room hay más variación en general. En el caso de hotel y shared room la mayoría tienen buen host response y acceptance rate. Es importante considerar que los 0 son imputaciones hechas para los valores nulos.

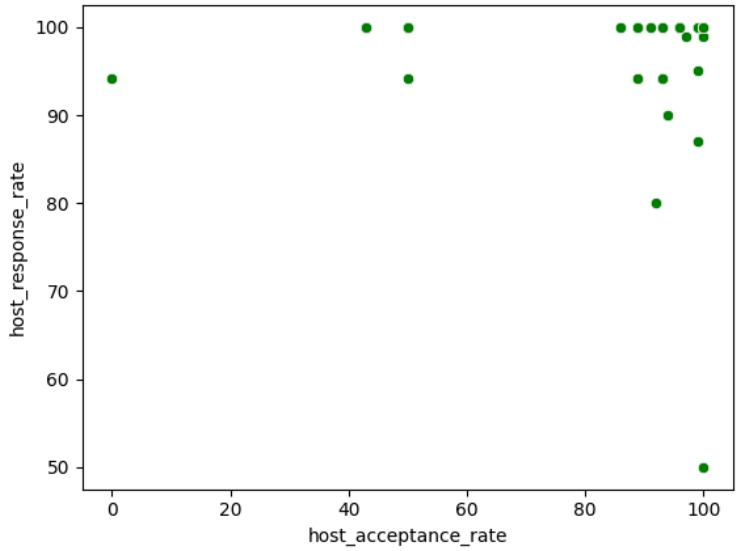
Entire home:



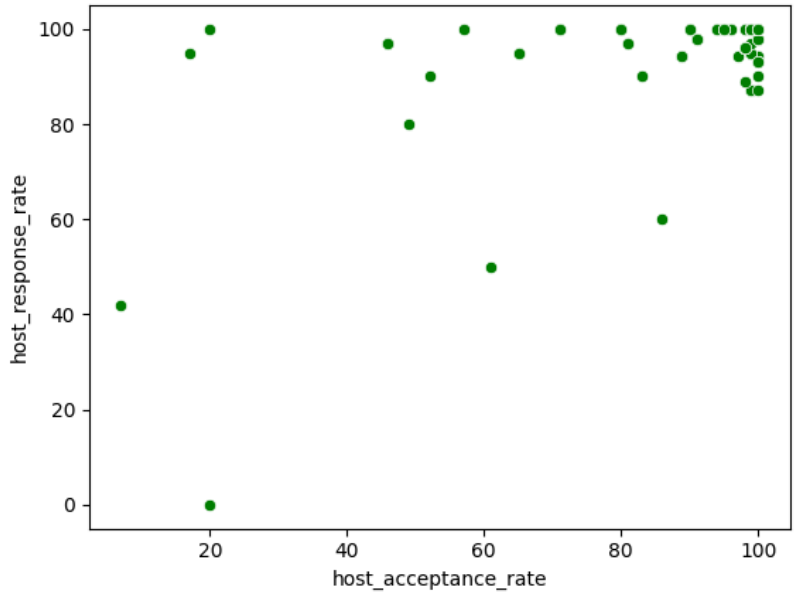
Private room:



Hotel:



Shared room:

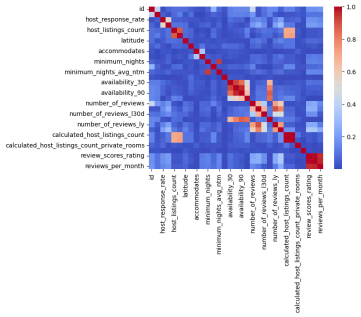


5)

Entire home:

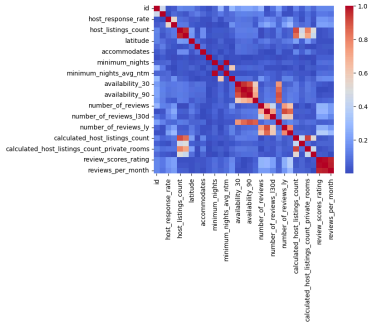
Calculated host listings count vs calculated host listings count entire homes	0.99
Review scores rating vs review score communication	0.98
Availability 60 vs availability 90	0.97

Review scores communication vs reviews per month	0.96
Review scores rating vs reviews per month	0.96
Minimum nights vs minimum nights avg ntm	0.93
Availability 30 vs availability 60	0.92
Number of reviews ltm vs number of reviews ly	0.90
Host listings count vs host total listings count	0.89
Availability 90 vs availability eoy	0.87



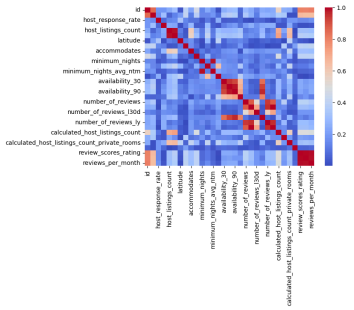
Private room

Minimum nights vs ,minimum nights atm	0.99
Review scores rating vs review scores communication	0.99
Availability 60 vs availability 90	0.97
Review scores communication vs reviews per month	0.97
Review scores rating vs reviews per month	0.96
Host listings count vs host total listings count	0.96
Availability 30 vs availability 60	0.94
Calculated host listings count vs calculated host listings count private rooms	0.92
Number of reviews ltm vs number of reviews ly	0.91
Host listings count vs calculated host listings count	0.90



Hotel:

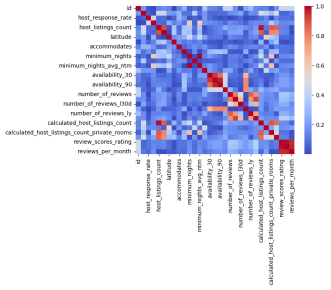
Number of reviews ltm vs estimated occupancy 365	1
Review scores rating vs review scores communication	0.99
Host listings count vs host total listings count	0.98
Minimum nights vs minimum nights avg ntm	0.98
Review scores communication vs reviews per month	0.98
Availability 60 vs availability 90	0.98
Availability 90vs availability eoy	0.96
Number of reviews ltm vs number of reviews ly	0.96
Number of reviews ly vs estimated occupancy l365	0.96



Shared room

Minimum nights vs minimum nights avg ntm	0.99

Host listings count vs calculated host listings count	0.98
Review scores rating vs review scores communication	0.98
Availability 60 vs availability 90	0.97
Review scores communication vs reviews per month	0.96
Review scores rating vs reviews per month	0.95
Availability 30 vs availability 60	0.93
Number of reviews ltm vs estimated occupancy l365	0.92
Host listings count vs host total listings count	0.90
Calculated host listings count vs calculated host listings count pr	0.86

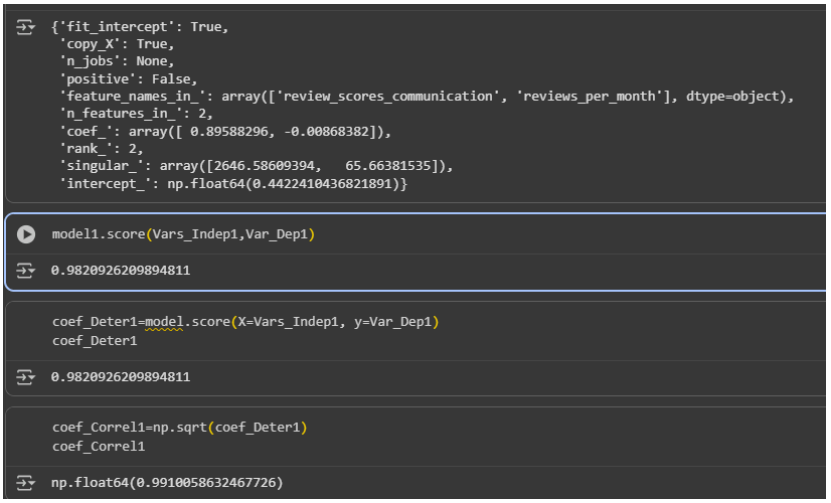


De este paso puedo concluir que las variables con mayor correlación son similares en los 4 tipos de alojamientos. Algunas que me parecieron interesantes fueron reviews score rating vs review scores communication ya que muestra la importancia de la atención a l@s huéspedes en la calificación final, review scores communication vs reviews per month ya que nuevamente se muestra que cuando hay buen servicio se realizan más reservas y review scores rating vs reviews per month ya que indica que mientras más reservas el anfitrión mejora y tiene mejores reseñas.

6)

Nota: para identificar las correlaciones de las regresiones simples estuve usando la información de entire home debido a que es el tipo de cuarto más común, sin embargo,varía en cada uno de ellos, lo que afecta el resultado final.

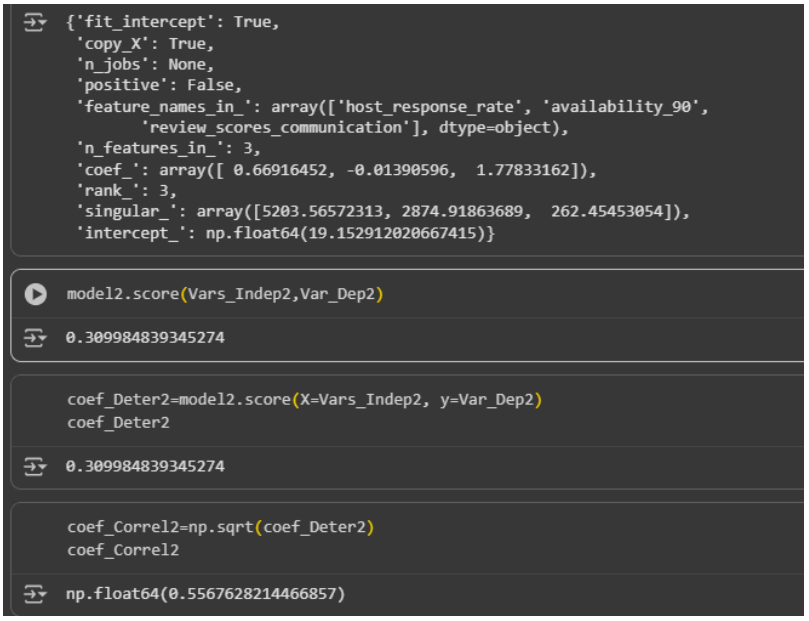
Review scores rating



Modelo: y = 0.89x1 -0.0006x2 + 0.44

Este modelo tiene un coeficiente de determinación de 0.98 y de correlación de 0.99, por lo que es muy confiable para predecir review scores rating, es decir la calificación más relevante y que resume el desempeño del anfitrión.

Host acceptance rate



Modelo: y = 0.66x1 -0.013x2 + 1.77x3 +19.15

Predecir la tasa de aceptación que tienen los anfitriones es difícil debido a las bajas correlaciones que existen con esta variable, sin embargo, al combinar host acceptance rate, availability 90 y review scores communication se alcanza 0.3 en determinación y .56 en correlación.

Host is superhost

Modelo: y= 0.001x1 + 0.003x2 + 0.002x3 -0.16

La variable si host es superhost no cuenta con muy altas correlaciones, sin embargo, al aplicar las más altas (estimated occupancy l365d, host acceptance rate y host response rate) se puede subir la determinación a 0.148 y la correlación a 0.38.

Host total listings count

Modelo: y= 1.06x1 + 0.63x2 -2.34

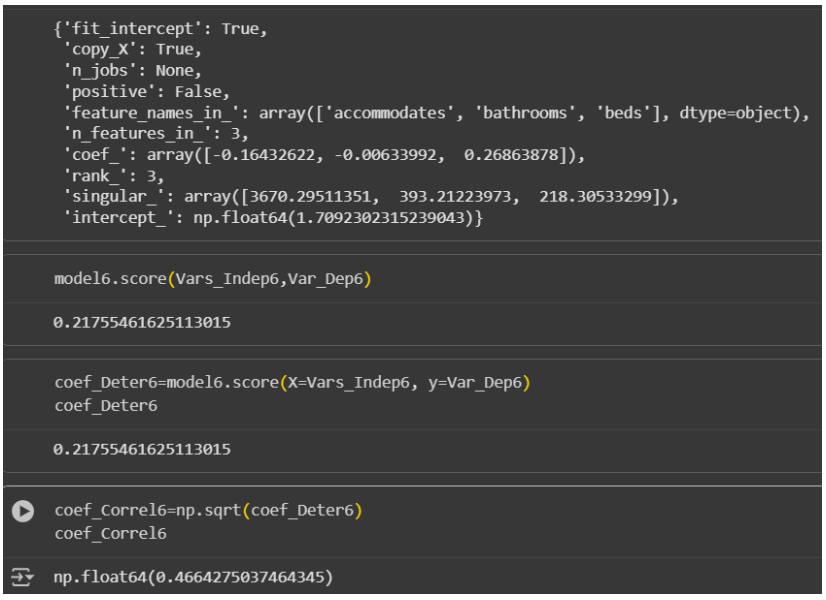
Esta variable ya contaba con coeficientes altos, que todavía aumentan gracias al modelo múltiple con las variables independientes host listings count y calculated host listings count, la determinación y correlación son 0.88 y 0.89 respectivamente.

Accommodates

Modelo: y= -2.46x1 +2.60x2 + 3.11x3 +2.49

Pocas variables están muy correlacionadas con accommodates, sin embargo escogí las más altas (bedrooms, estimated revenue l365d y price) para obtener los coeficientes de 0.07 y 0.27 respectivamente. Algo interesante es que cuando aumentan los cuartos se resta en el modelo, disminuyendo accommodates.

Bedrooms



Modelo: y= -0.16x1 -0.006x2 + 0.27x3 + 1.7

Aunque en general la correlación con bedrooms no es muy alta, combinar accommodates, bathrooms y beds que son las más relevantes (y que además se relacionan de forma lógica) lleva a .22 de determinación y 0.47 de correlación. Algo interesante es que tanto accommodates como bathrooms restan en el modelo, cuando se podría considerar al revés.

Price

Modelo: y= 7.54x1 + 1.82x2 - 3.64x3 + 168.4

A pesar de haber combinado las variables más correlacionadas, los coeficientes para price siguen siendo muy bajos (0.01 y 0.11 respectivamente). Considero que esto se debe a que en otros tipos de cuarto hay variables más relevantes que estas, es decir que el dataset varía considerablemente.

Review scores value

Modelo: y= 0.59x1 + 0.31x2 +0..09x3 -0.0005

La variable de review scores value se relaciona fuertemente con otras reviews, al combinarla se pueden alcanzar coeficientes casi perfectos de 0.99 tanto para determinación como para correlación.

Bathrooms

```
{'fit_intercept': True,
 'copy_X': True,
 'n_jobs': None,
 'positive': False,
 'feature_names_in_': array(['beds', 'availability_90', 'availability_365'], dtype=object),
 'n_features_in_': 3,
 'coef_': array([ 9.92766646e-01, -3.58842462e-04, -8.79541810e-04]),
 'rank_': 3,
 'singular_': array([20532.39170053,  3879.21275364,  2222.82579997]),
 'intercept_': np.float64(-0.18460978820628426)}
```

```
model9.score(Vars_Indep9,Var_Dep9)
```

```
0.9830014211215022
```

```
coef_Deter9=model9.score(X=Vars_Indep9, y=Var_Dep9)
coef_Deter9
```

```
0.9830014211215022
```

```
coef_Correl9=np.sqrt(coef_Deter9)
coef_Correl9
```

```
np.float64(0.9914642813140079)
```

Modelo: y= 9.93x1 - 3.59x2 -8.79x3 -0.18

El modelo múltiple para bathrooms también es muy bueno, ya que al combinar variables que tienen alta correlación (beds, availability 90 y availability 365) se obtienen coeficientes de 0.98 para determinación y 0.99 para correlación. Las variables de availability me llaman la atención ya que cuando un alojamiento tiene más disponibilidad disminuye el número de baños con los que cuenta.

Reviews per month

```
{'fit_intercept': True,
 'copy_X': True,
 'n_jobs': None,
 'positive': False,
 'feature_names_in_': array(['review_scores_checkin', 'review_scores_communication',
       'review_scores_location'], dtype=object),
 'n_features_in_': 3,
 'coef_': array([-2.77654914, -0.78885786, -6.12083891]),
 'rank_': 3,
 'singular_': array([460.58391056,  33.28280758,  23.79749508]),
 'intercept_': np.float64(48.83672540905934)}
```

```
model10.score(Vars_Indep10,Var_Dep10)
```

```
0.9562957324386572
```

```
coef_Deter10=model10.score(X=Vars_Indep10, y=Var_Dep10)
coef_Deter10
```

```
0.9562957324386572
```

```
coef_Correl10=np.sqrt(coef_Deter10)
coef_Correl10
```

```
np.float64(0.9779037439536967)
```

Modelo: y= -2.78x1 -0.79x2 -6.12x3 +48.84

Reviews per month está muy correlacionada con las distintas reviews de forma negativa, es decir que cuando aumenta la calificación disminuye el número de reviews. El modelo múltiple es muy bueno, aunque mejora poco respecto a los simples, su determinación es de 0.96 y su correlación de 0.98.

Referencia del dataset:

Get the Data. (s.f.). Inside Airbnb. https://insideairbnb.com/get-the-data/