PROYECTO FINAL

Análisis AirBnb Madrid 2012-2016

Equipo de proyecto

loop
Infinite loop

Susana Camacho
Erica Martínez
Fátima Ramírez
Josselyn Maritza Jumpa
Cristina Porta

ÍNDICE

ANÁLISIS AIRBNB 2012-2016

ÍNDICE

ANÁLISIS AIRBNB 2012-2016

OBJETIVOS

En el año 2019 AirBnB crea un nuevo proyecto para aumentar su cuota de mercado en capitales europeas. En la división del sur de Europa, el equipo de analistas **Infinite loop** realiza un estudio exhaustivo de las propiedades listadas ubicadas en la ciudad de Madrid para determinar si las reseñas de las propiedades influyen en su precio.

ANÁLISIS EXPLORATORIO

ANÁLISIS AIRBNB 2012-2016

PRE-PROCESAMIENTO DE DATOS

- Revisión de valores nulos en la tabla: eliminación o imputación.
- Limpieza de columnas: eliminación, encoding o limpieza de texto.

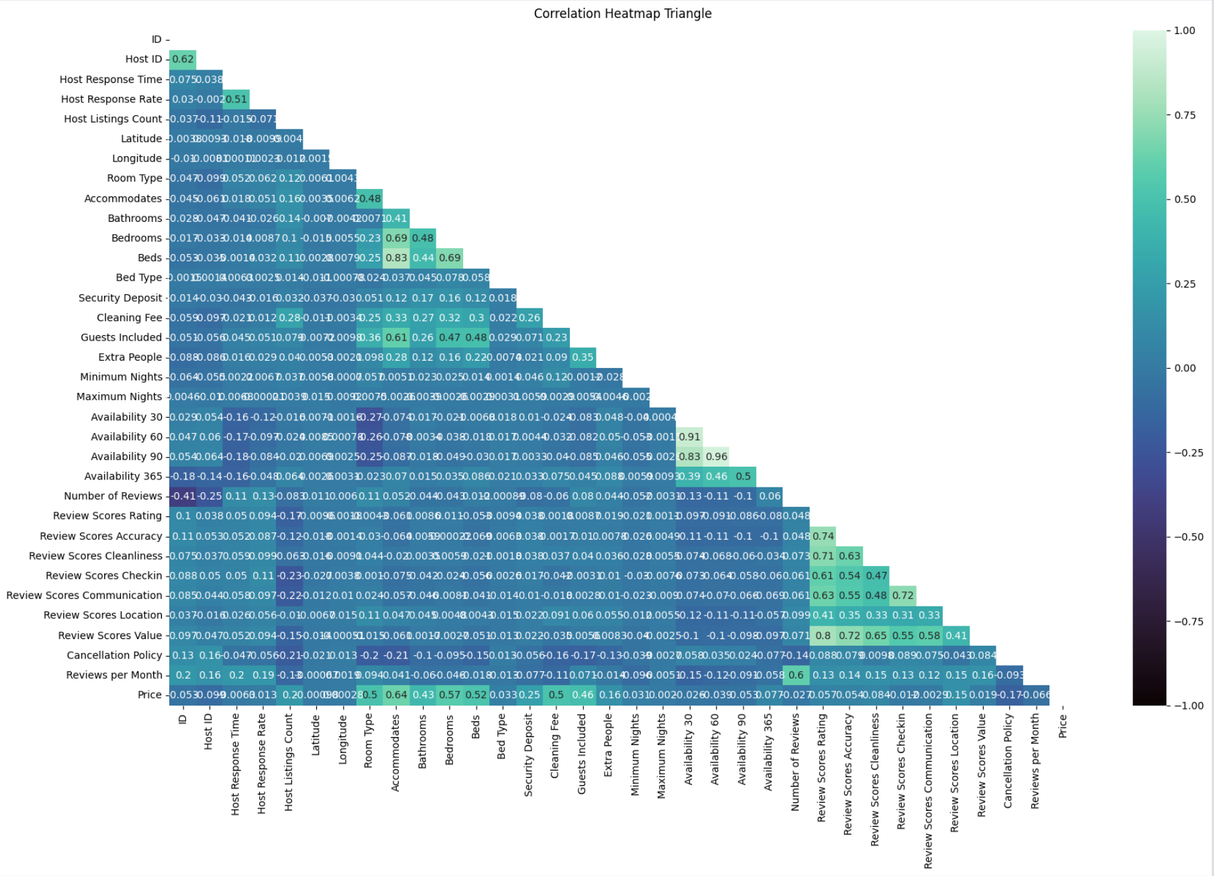
VISUALIZACIÓN DE DATOS

- Detección de outliers con boxplots y limpieza de éstos.
- Comparación de dataframes iniciales y limpios.
- Revisión de altas correlaciones entre columnas.
- Palabras clave.

Nube de palabras de las descripciones de las propiedades.

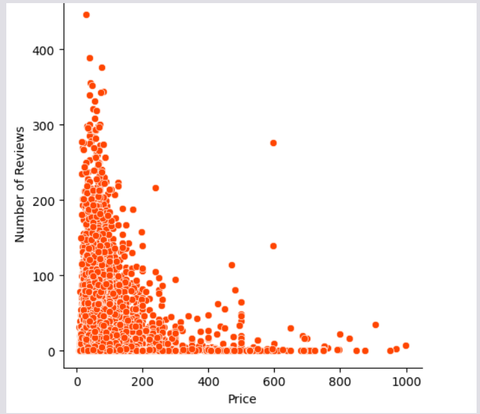
ANÁLISIS EXPLORATORIO

ANÁLISIS AIRBNB 2012-2016

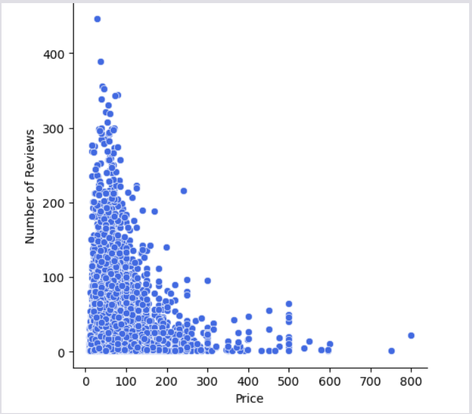


Mapa de calor de las correlaciones entre columnas.

Dataframe original



Dataframe con valores nulos de reviews imputados

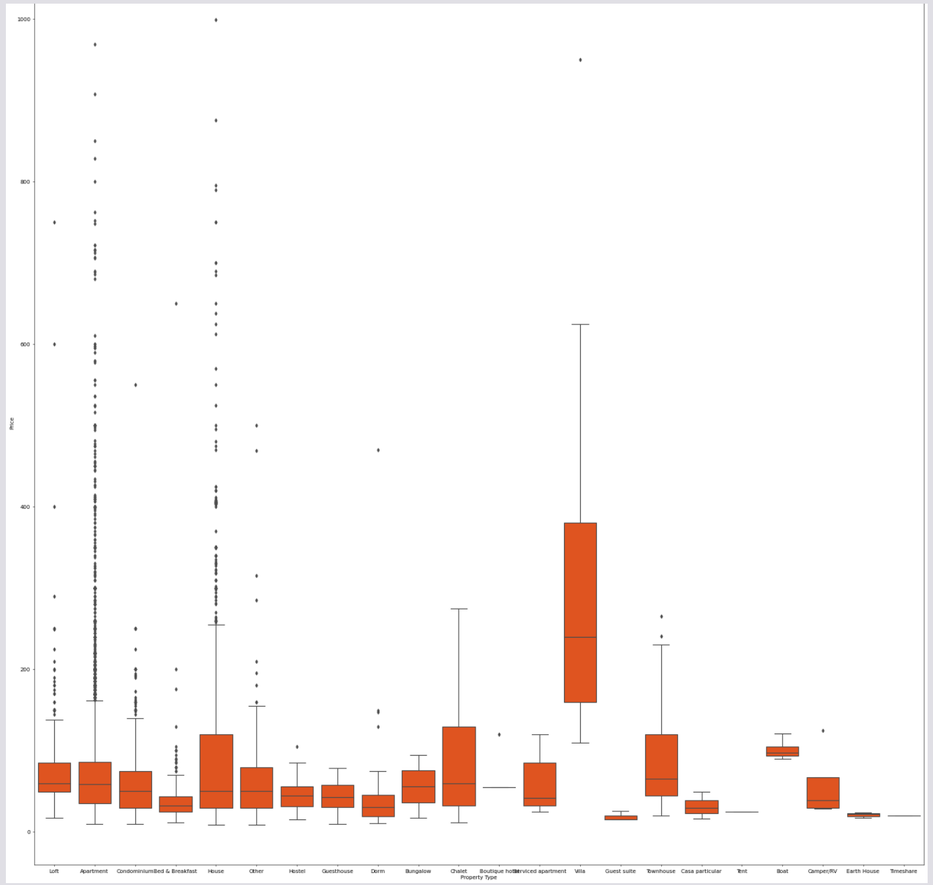


Distribución de las reviews según precio.

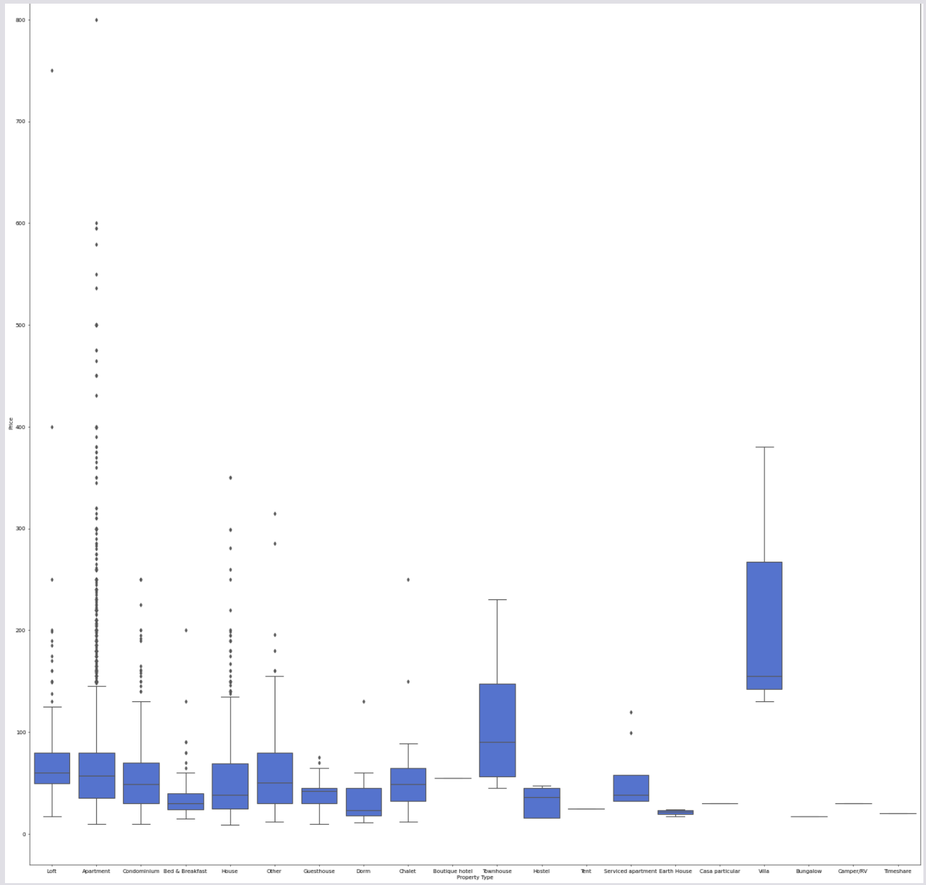
ANÁLISIS EXPLORATORIO

ANÁLISIS AIRBNB 2012-2016

Dataframe original



Dataframe con valores nulos de reviews imputados

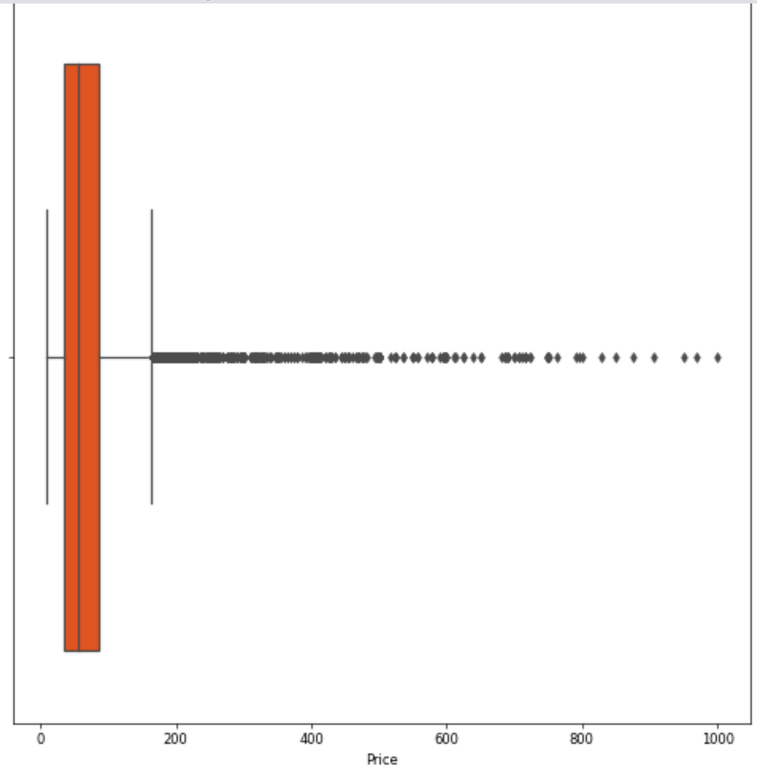


Distribución del precio según tipo de alojamiento.

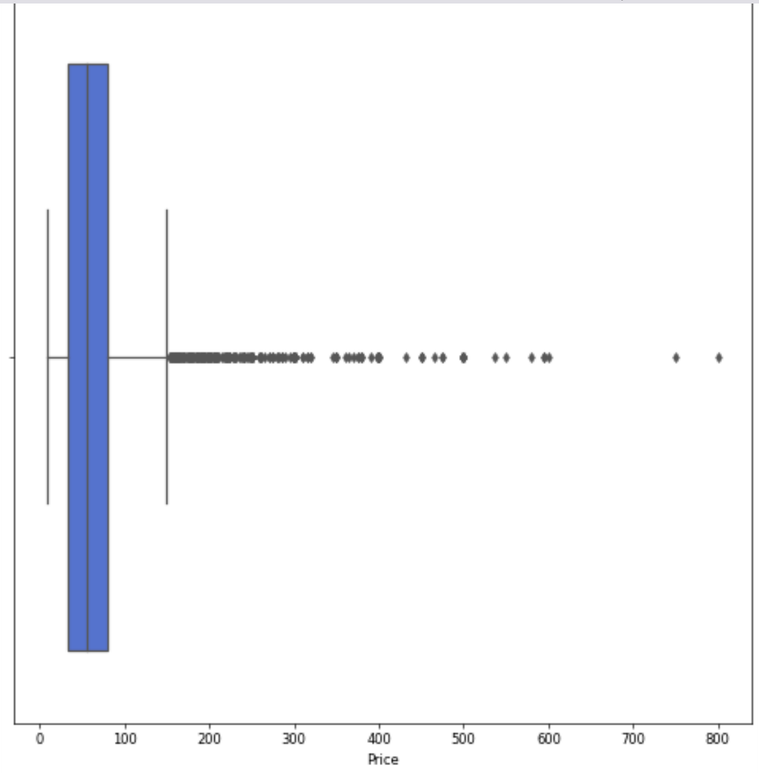
ANÁLISIS EXPLORATORIO

ANÁLISIS AIRBNB 2012-2016

Dataframe original



Dataframe con valores nulos de reviews imputados



Distribución del precio.

ARQUITECTURA Y VALIDACIÓN DE LOS DATOS

REQUERIMIENTOS PARA CONECTAR PYTHON CON POSTGRES

- **La librería psycopg2** permite conectarse y trabajar con bases de datos PostgreSQL.
- Los credenciales necesarios para conectarse a la base de datos: ***hostname, password, port_ID, database name***
- La función **connect** inicia la conexión con postgres para iniciar cualquier transacción DDL o DML, y **close** la cierra.
- La función **cursor** permite ejecutar los comandos en el contexto de nuestra sesión de Postgres.
- La función **execute** permite ejecutar el comando que le demos, por ejemplo: crear la base de datos airbnb, crea una tabla, inserta datos en la tabla.
- La conexión con Posgres se ha controlado mediante el manejo de errores (Exception handling: ***try, except, finally***) para configurar el error en el caso de que la conexión falle.

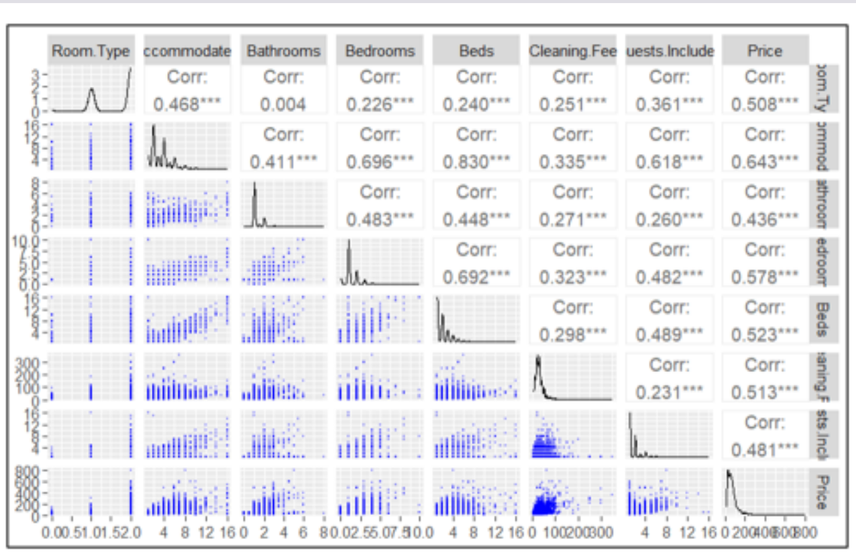
```
### Connect to Postgres and create 'airbnb' database

try:
    dbConnection = psycopg2.connect(
        user = "postgres",
        password = "***",
        host = "localhost",
        port = "5433",
        database = "postgres")
    dbConnection.set_isolation_level(0) # AUTOCOMMIT
    dbCursor = dbConnection.cursor()
    dbCursor.execute('CREATE DATABASE airbnb;')
    dbCursor.close()
except (Exception , psycopg2.Error) as dbError :
    print ("Error while connecting to PostgreSQL", dbError)
finally:
    if(dbConnection): dbConnection.close()
```

MODELO DE REGRESIÓN LINEAL

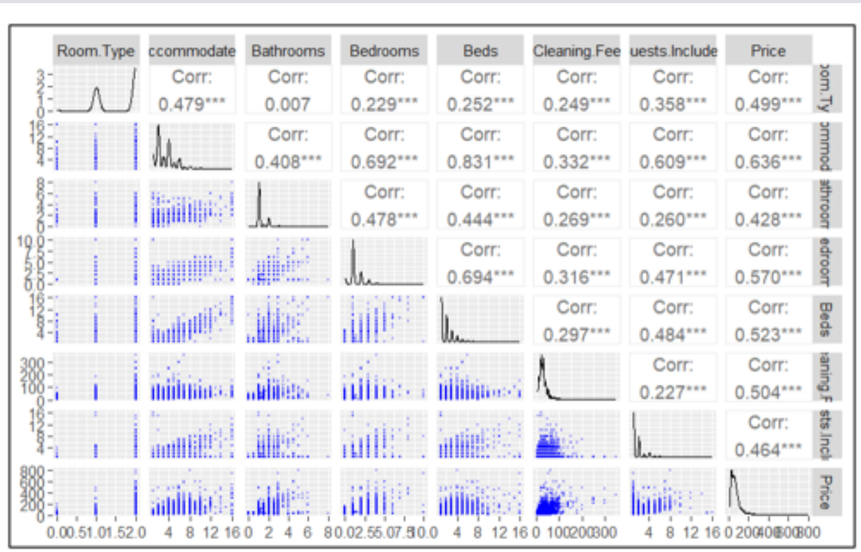
ANÁLISIS AIRBNB 2012-2016

1.ANÁLISIS DE CORRELACIÓN

- **Comparar resultados entre dos datasets:** NA reviews deleted vs NA imputados.
- Ninguna de las variables explicativas sigue una distribución normal.
- La relación entre el precio y las variables independientes es positiva



Dataframe no reviews



DF reviews imputed

MODELO DE REGRESIÓN LINEAL

ANÁLISIS AIRBNB 2012-2016

2. COMPARACIÓN DE LOS MODELOS

Model	R-squared	Adjusted R^2	RSE
Model 1	0.62	0.619	30.88
Model 2	0.6383	0.6368	30.17
Model 3	0.6373	0.6364	30.18
Model 4	0.6337	0.633	30.33
Model 5	**0.618**	**0.6175**	**30.96**
Model 6	0.6147	0.6143	61.09

Tabla 1: Resumen del R cuadrado, R ajustado y RSE *(no reviews)*

Model	R-squared	Adjusted R^2	RSE
Model 1	0.6032	0.6026	31.63
Model 2	0.6199	0.6184	31
Model 3	0.6145	0.6139	31.18
Model 4	0.6093	0.6088	31.38
Model 5	**0.6028**	**0.6024**	**31.64**
Model 6	0.6	0.5996	31.75

Tabla 2: Resumen del R cuadrado, R ajustado y RSE *(reviews imputed)*

```
summary(model5)


Call:
lm(formula = Price ~ . - Host.Response.Time - Host.Response.Rate -
    Beds - Bed.Type - Maximum.Nights - Availability.365 - Number.of.Reviews -
    Review.Scores.Accuracy - Review.Scores.Checkin - Review.Scores.Value -
    Cancellation.Policy - Minimum.Nights - Review.Scores.Cleanliness -
    Review.Scores.Communication - Host.Listings.Count - Extra.People -
    Review.Scores.Rating - Review.Scores.Location - Reviews.per.Month,
    data = airbnb.train)

Residuals:
    Min      1Q  Median      3Q     Max
-199.16  -13.99   -1.68   11.00  545.00

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -73.321115   1.922723 -38.134  < 2e-16 ***
Room.Type         32.114103   0.943488  34.038  < 2e-16 ***
Accommodates       3.740438   0.324267  11.535  < 2e-16 ***
Bathrooms         16.500068   0.816830  20.200  < 2e-16 ***
Bedrooms          10.263089   0.678752  15.121  < 2e-16 ***
Security.Deposit   0.066480   0.005248  12.667  < 2e-16 ***
Cleaning.Fee       0.637541   0.023499  27.131  < 2e-16 ***
Guests.Included    3.207081   0.425421   7.539 5.39e-14 ***
Availability.30    0.481120   0.050672   9.495  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.96 on 6567 degrees of freedom
Multiple R-squared:  0.618,     Adjusted R-squared:  0.6175
F-statistic:  1328 on 8 and 6567 DF,  p-value: < 2.2e-16
```

Dataframe no reviews

```
summary(model11)


Call:
lm(formula = Price ~ . - Host.Response.Time - Host.Response.Rate -
    Beds - Bed.Type - Maximum.Nights - Availability.365 - Number.of.Reviews -
    Review.Scores.Accuracy - Review.Scores.Checkin - Review.Scores.Value -
    Cancellation.Policy - Minimum.Nights - Review.Scores.Cleanliness -
    Review.Scores.Communication - Host.Listings.Count - Extra.People -
    Review.Scores.Rating - Review.Scores.Location - Reviews.per.Month -
    Reviews.per.Month, data = airbnb2.train)

Residuals:
    Min      1Q  Median      3Q     Max
-196.94  -14.31   -1.71   10.93  478.41

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -69.730577   1.856997 -37.550  < 2e-16 ***
Room.Type         30.855880   0.910280  33.897  < 2e-16 ***
Accommodates       4.142097   0.311070  13.316  < 2e-16 ***
Bathrooms         15.051778   0.815929  18.447  < 2e-16 ***
Bedrooms          10.768474   0.651580  16.527  < 2e-16 ***
Security.Deposit   0.061057   0.005093  11.988  < 2e-16 ***
Cleaning.Fee       0.625388   0.022167  28.212  < 2e-16 ***
Guests.Included    2.917326   0.412426   7.074 1.65e-12 ***
Availability.30    0.493447   0.047443  10.401  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.64 on 7117 degrees of freedom
Multiple R-squared:  0.6028,    Adjusted R-squared:  0.6024
F-statistic:  1350 on 8 and 7117 DF,  p-value: < 2.2e-16
```

DF reviews imputed

$$\hat{Y} = -73.3 + 32.1(\text{Room.Type}) + 3.7(\text{Accommodates}) + 16.5(\text{Bathrooms}) + 10.2(\text{Bedrooms}) + 0.06(\text{Security.Deposit}) + 0.6(\text{Cleaning.Fee}) + 3.2(\text{Guests.Included}) + 0.48(\text{Availability30})$$

MODELO DE REGRESIÓN LINEAL

ANÁLISIS AIRBNB 2012-2016

3. CALIDAD DEL MODELO

```r
airbnb.train$price_est <- predict(model5, airbnb.train)
caret::postResample(pred = airbnb.train$price_est, obs=airbnb.train$Price)
```

```
      RMSE   Rsquared        MAE
30.9359210  0.6180093 18.3615774
```

```r
airbnb.test$price_est <- predict(model5, airbnb.test)
caret::postResample(pred = airbnb.test$price_est, obs=airbnb.test$Price)
```

```
      RMSE   Rsquared        MAE
28.9472443  0.6432131 18.9317040
```

Dataframe no reviews

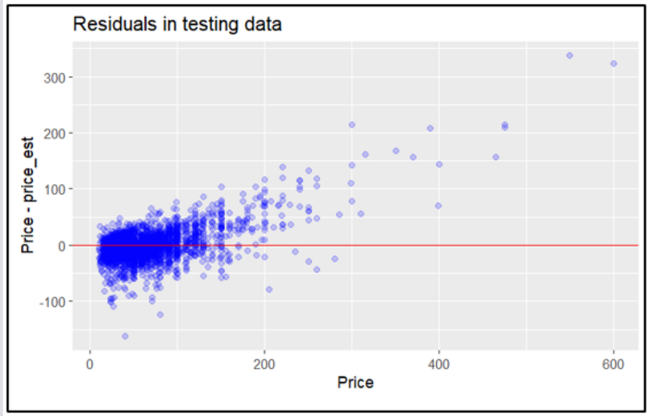
```r
airbnb2.train$price_est2 <- predict(model11, airbnb2.train)
caret::postResample(pred = airbnb2.train$price_est2, obs=airbnb2.train$Price)
```

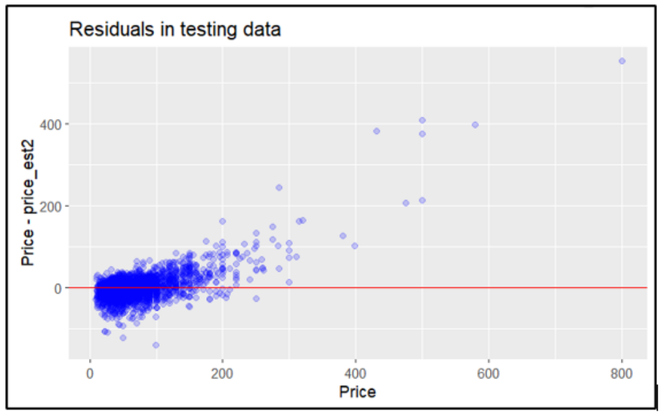
```
      RMSE   Rsquared        MAE
31.6214913  0.6028347 18.7519209
```

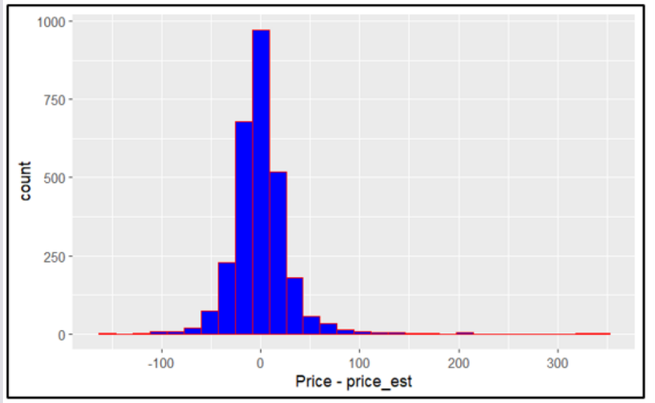
```r
airbnb2.test$price_est2 <- predict(model11, airbnb2.test)
caret::postResample(pred = airbnb2.test$price_est2, obs=airbnb2.test$Price)
```

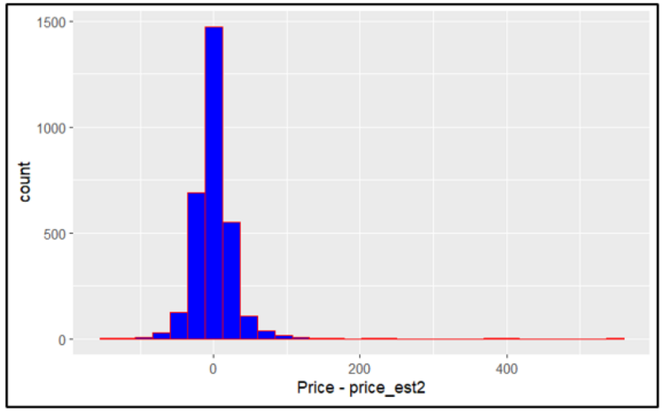
```
      RMSE   Rsquared        MAE
31.0061525  0.6181757 18.1799185
```

DF reviews imputed









Dataframe no reviews

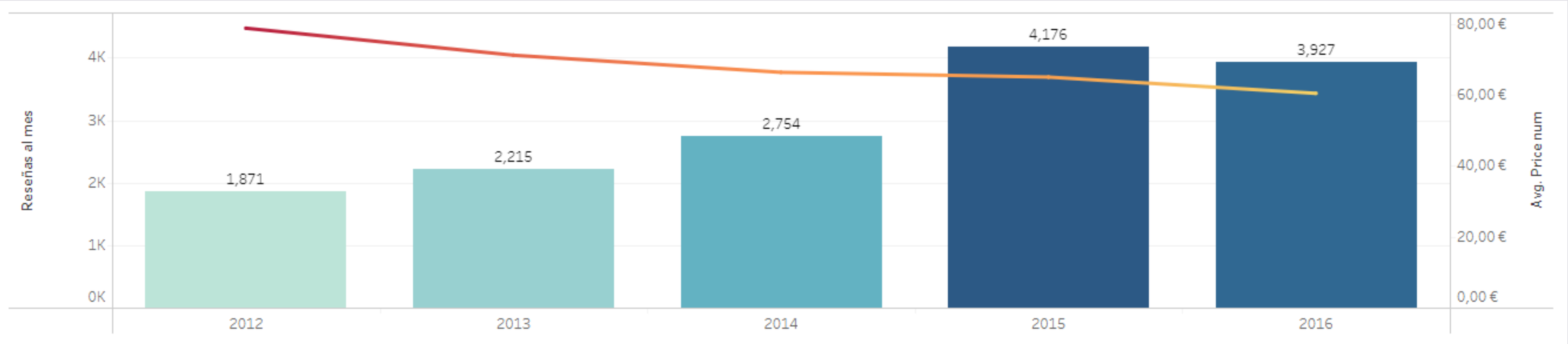
DF reviews imputed

VISUALIZACIÓN DE LAS METRICAS

ANÁLISIS AIRBNB 2012-2016

DASHBOARD

https://public.tableau.com/app/profile/josselyn.jumpa/viz/Airbnb_visual/Dashboard



GRACIAS

POR VUESTRA ATENCIÓN

loop

Infinite loop