APC
CAS KAGGLE

ENRIC FERRERA - 1672969
POL VILA - 1674294

Índex

01 **Introducció**

Introducció del cas Kaggle

02 **EDA (Exploratory Data Analysis**

Anàlisis i exploració del dataset

03 **Model i resultats**

04 **Conclusions**

01

Introducció

Introducció del cas Kaggle

DATASET

- Atributs del dataset

- id: identificador
- name: nom de l'anunci
- host_id: id del host
- neighbourhood_group: localització
- neighbourhood: barri
- latitude: latitud
- longitude: longitud
- room_type: tipus d'apartament
- price: preu en dolars (target)
- number_of_reviews: número de reviews
- last_review: data de l'ultima ressenya
- calculated_host_listings_count: número de apartamentos que tiene el anfitrion
- availability_365: número de dias que está disponible para reservar

✓ **10 NUMÈRICS**

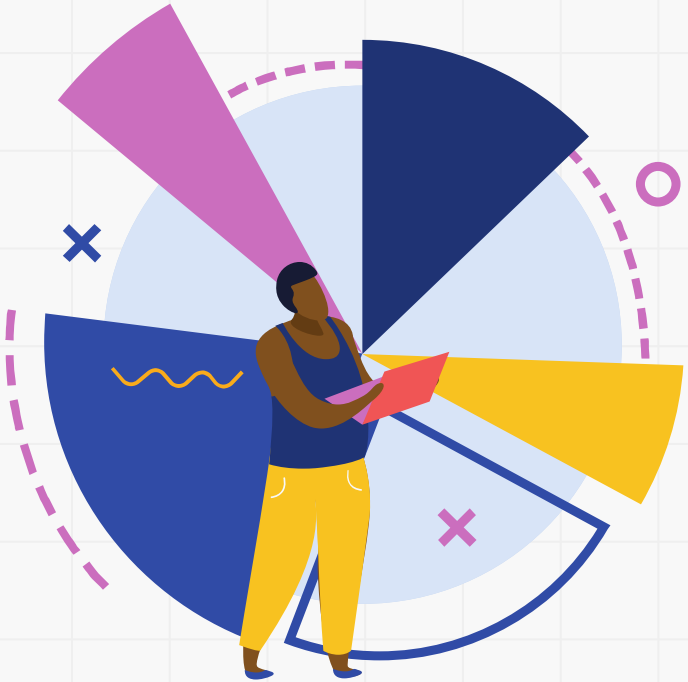
✓ **6 CATEGÒRICS**

48895 mostres

02

EDA (Exploratory Data Analysis)

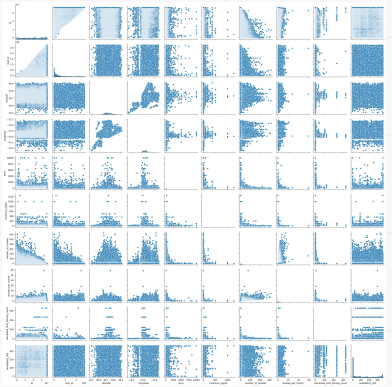
Anàlisis i exploració del dataset



Anàlisis descriptiu del dataset

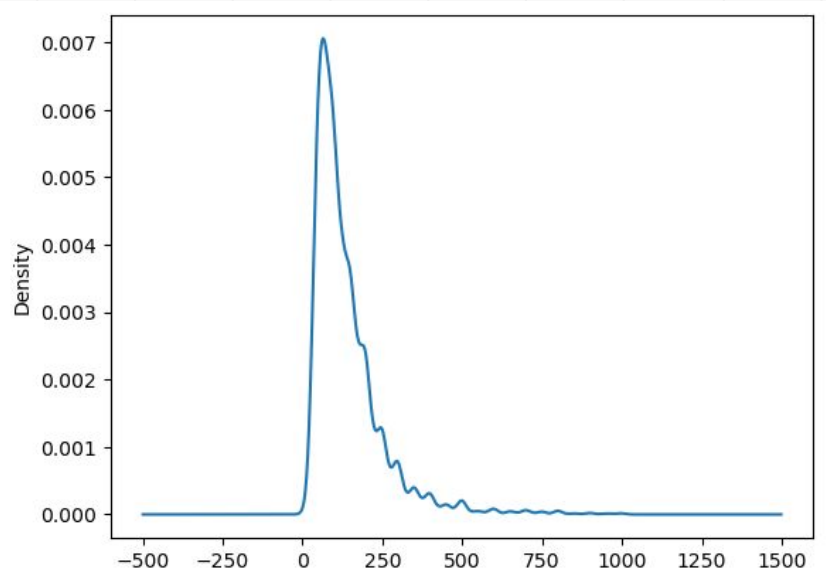
	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_c	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.00	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.14	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.95	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.00	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.00	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.00	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.00	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.00	365.000000

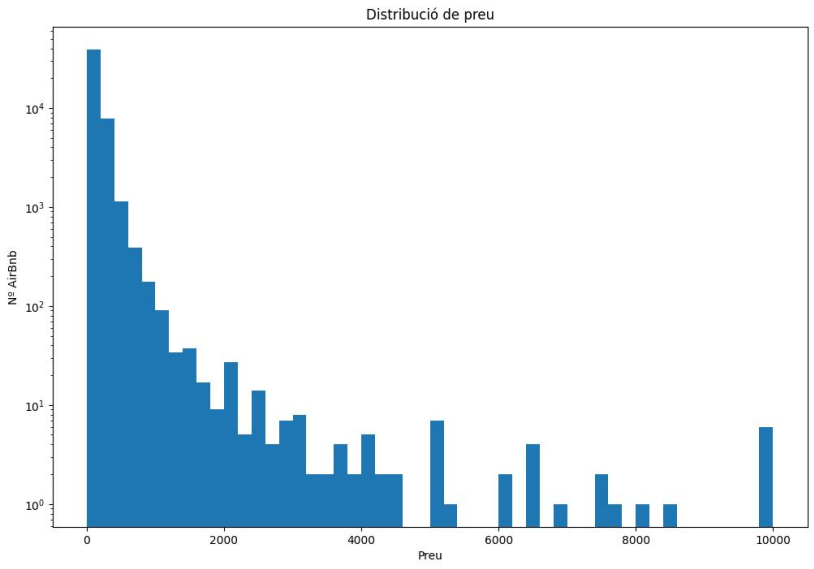
- ✓ La mitjana de preu ⇨ 152,72$
- ✓ Hi ha valors inconsitents



Distribució atribut target

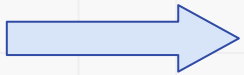
Gràfic de densitat





Preprocessament (valors nuls) i Data Cleaning

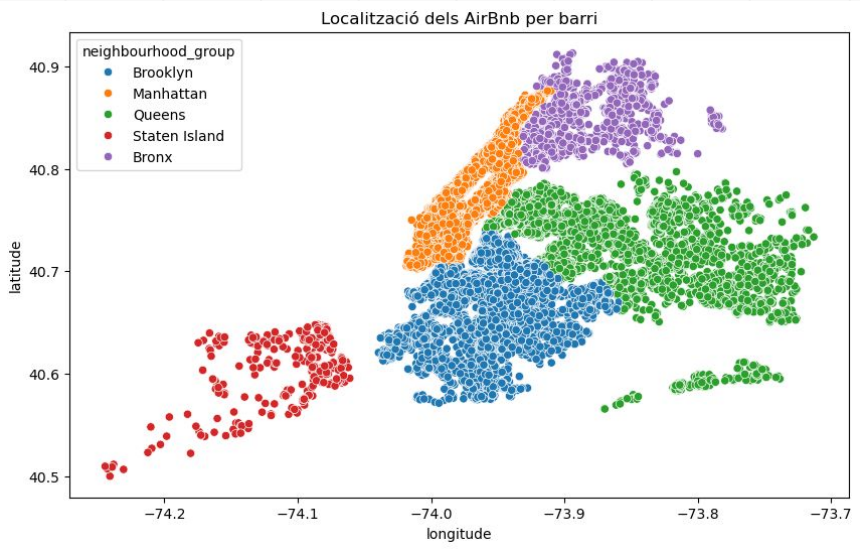
```
id                                0.000000
name                              0.000020
host_id                           0.000000
host_name                         0.000027
neighbourhood_group               0.000000
neighbourhood                     0.000000
latitude                          0.000000
longitude                         0.000000
room_type                         0.000000
price                             0.000000
minimum_nights                    0.000000
number_of_reviews                 0.000000
last_review                       0.012849
reviews_per_month                 0.012849
calculated_host_listings_count    0.000000
availability_365                  0.000000
```

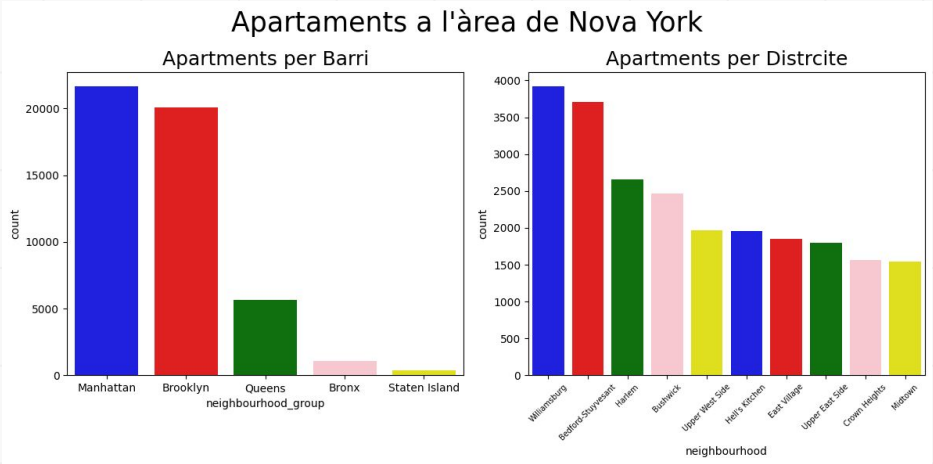


```
name
False    48879
True        16
Name: count, dtype: int64
host_name
False    48874
True        21
Name: count, dtype: int64
last_review
False    38843
True     10052
Name: count, dtype: int64
reviews_per_month
False    38843
True     10052
Name: count, dtype: int64
```

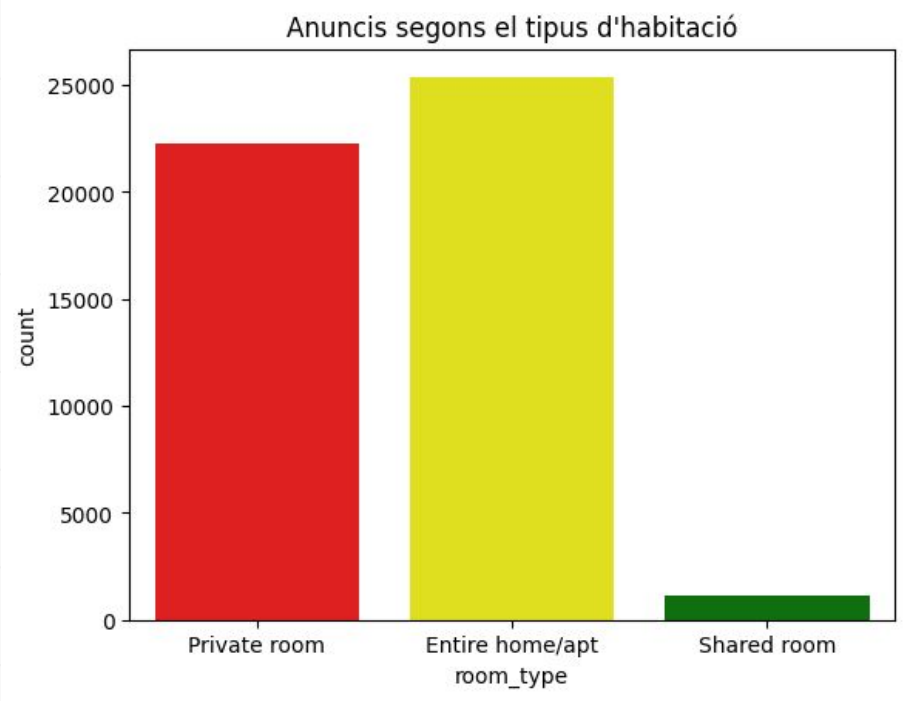
✓ Tractament dels valors nuls

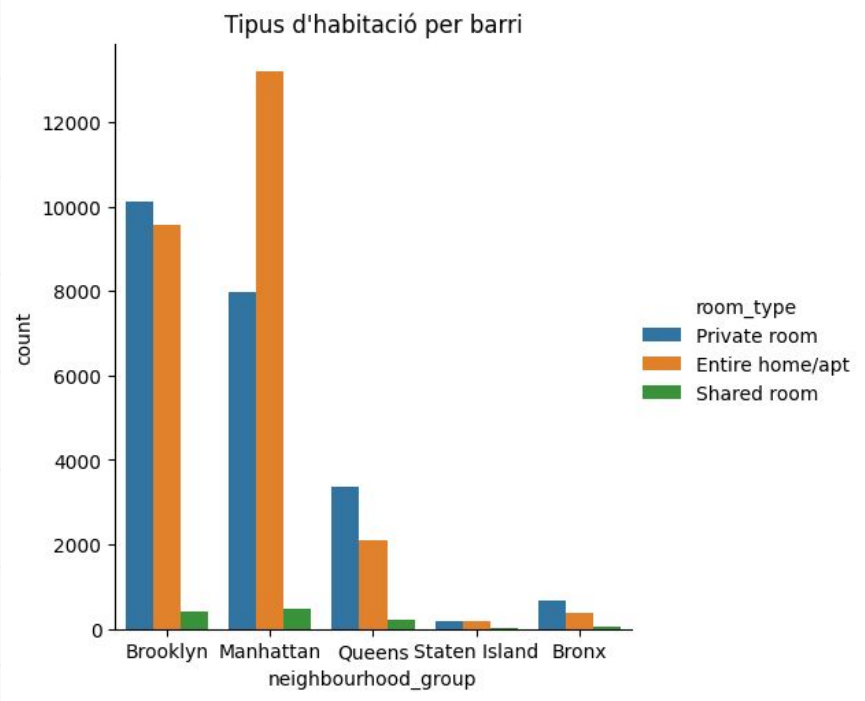
Distribució del anuncis



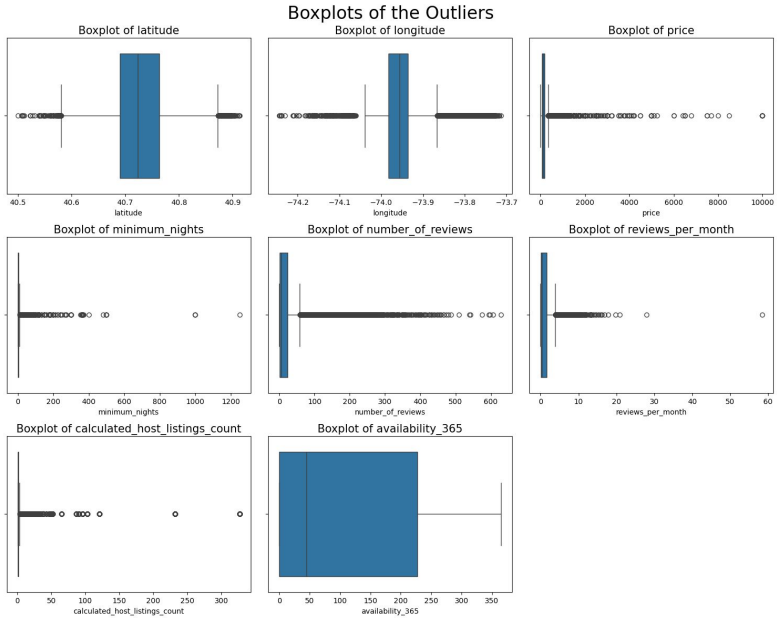


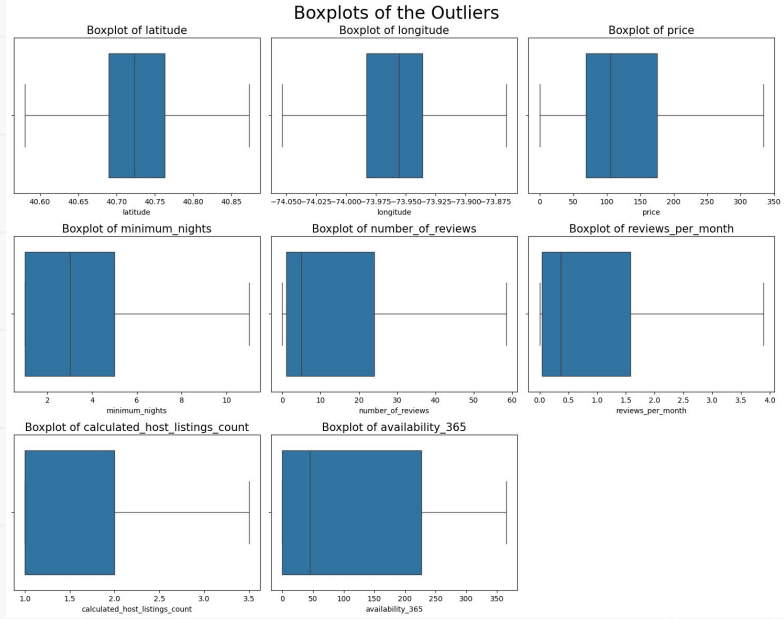
Anuncis per tipus d'habitació



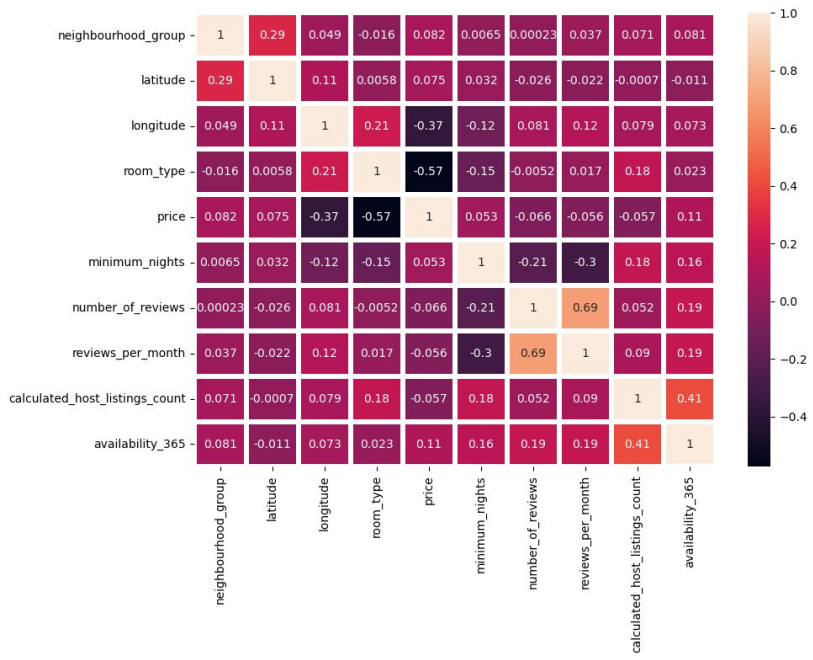


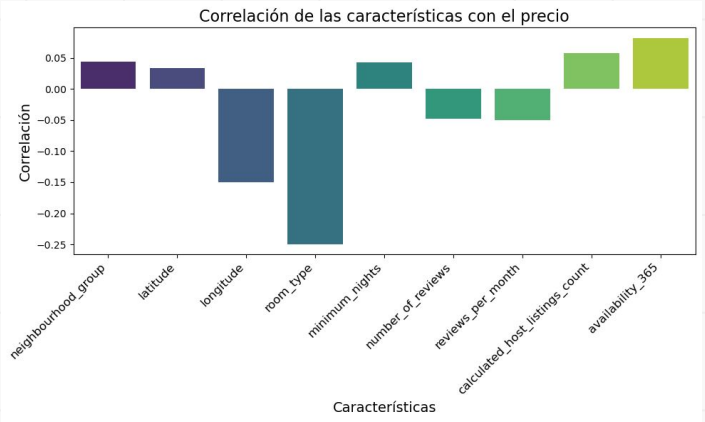
Outliers


Boxplots of the Outliers
Boxplot of latitude
Boxplot of longitude
Boxplot of price
Boxplot of minimum_nights
Boxplot of number_of_reviews
Boxplot of reviews_per_month
Boxplot of calculated_host_listings_count
Boxplot of availability_365
latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365

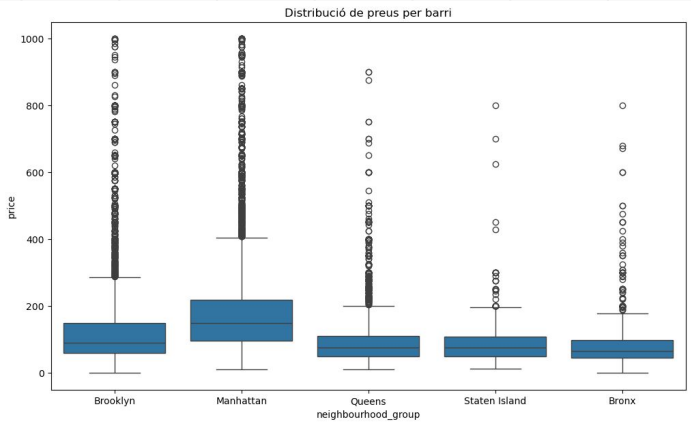

Boxplots of the Outliers
Boxplot of latitude
Boxplot of longitude
Boxplot of price
Boxplot of minimum_nights
Boxplot of number_of_reviews
Boxplot of reviews_per_month
Boxplot of calculated_host_listings_count
Boxplot of availability_365
latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365

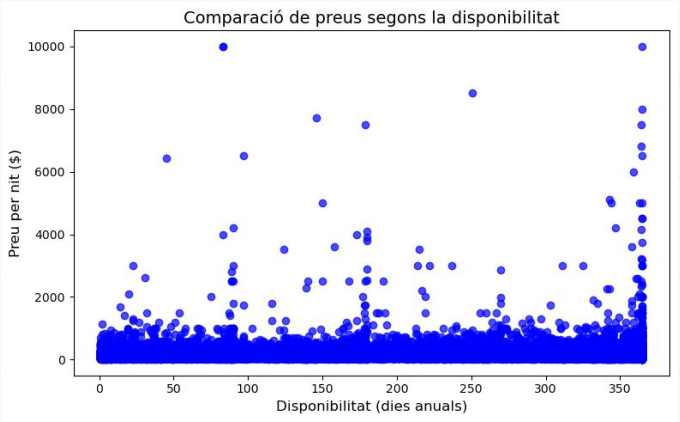
Matriu de correlació

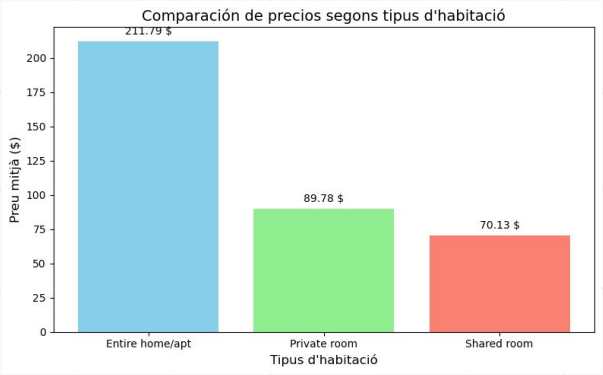


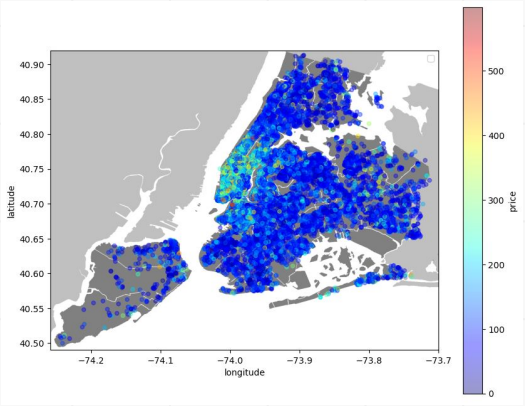


Relació variables amb atribut target









03

Model i resultats

Busqueda del millor model

Regressió lineal

```
0.44249208593895595
```

Regressió polinòmica

```
Degree: 2
R² en entrenamiento: 0.5156660838928089
R² en prueba: 0.5082671734532314
MSE en entrenamiento: 0.48433391610719095
MSE en prueba: 0.4917328265467686
```

Ridge

```
R² en entrenamiento: 0.5438318708444088
R² en prueba: 0.5240781975942029
```

Decision tree

```
R² en entrenamiento: 0.5781185341497777
R² en prueba: 0.5151369072669364
```

Random Forest

```
R² en entrenamiento: 0.9397364063650686
R² en prueba: 0.5514052979914996
```

Busquem els millors hiperparàmetres

```
Mejores parámetros encontrados: {'n_estimators': 500, 'min_samples_split': 10,
'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': 50}
```

```
R² en entrenamiento: 0.8208043024557491
R² en prueba: 0.5689185616589219
MAE: 39.160567997860625
MSE: 3015.6962105512375
```

Tractament dels outliers

Abans

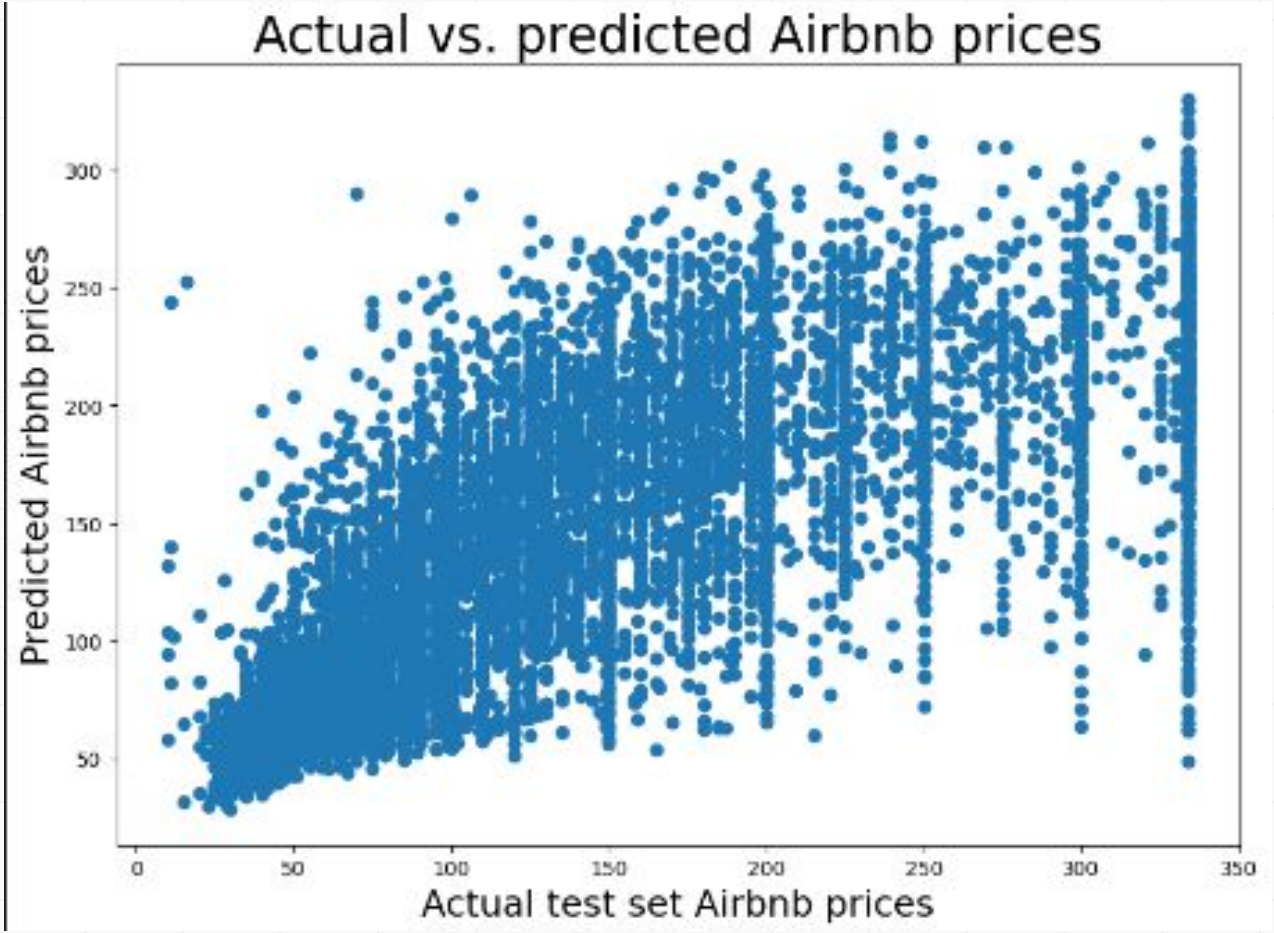
```
R² en entrenamiento: 0.8783039455680448
R² en prueba: 0.04170708417912328
MAE: 70.80684679152594
MSE: 52543.86322679357
```

Després

```
R² en entrenamiento: 0.8208043024557491
R² en prueba: 0.5689185616589219
MAE: 39.160567997860625
MSE: 3015.6962105512375
```

04

Conclusions


Actual vs. predicted Airbnb prices
Predicted Airbnb prices
Actual test set Airbnb prices
50
100
150
200
250
300
0
50
100
150
200
250
300
350