

# New York City AirBnB Data

---

Javier Méndez Leiva

[https://github.com/AuraKaz/APC-Kaggle-NewYork\\_City\\_Airbnb\\_Data](https://github.com/AuraKaz/APC-Kaggle-NewYork_City_Airbnb_Data)

# Introducción a la BD

id	int64
name	object
host_id	int64
host_name	object
neighbourhood_group	object
neighbourhood	object
latitude	float64
longitude	float64
room_type	object
price	int64
minimum_nights	int64
number_of_reviews	int64
last_review	object
reviews_per_month	float64
calculated_host_listings_count	int64
availability_365	int64

-16 columnas

-5 tipos textuales, 3 de ellos categóricos

Valores nulos

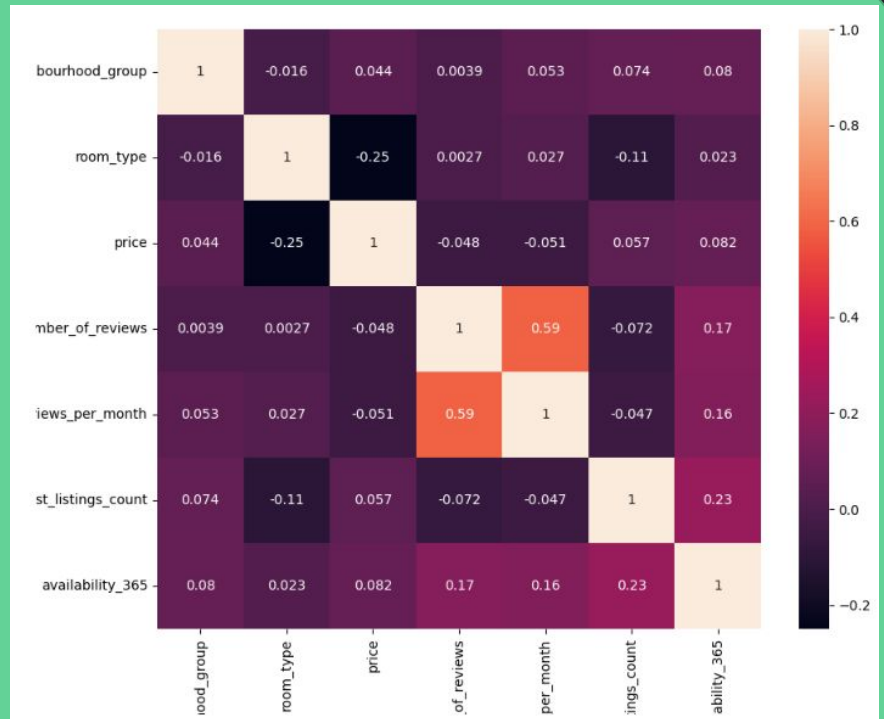
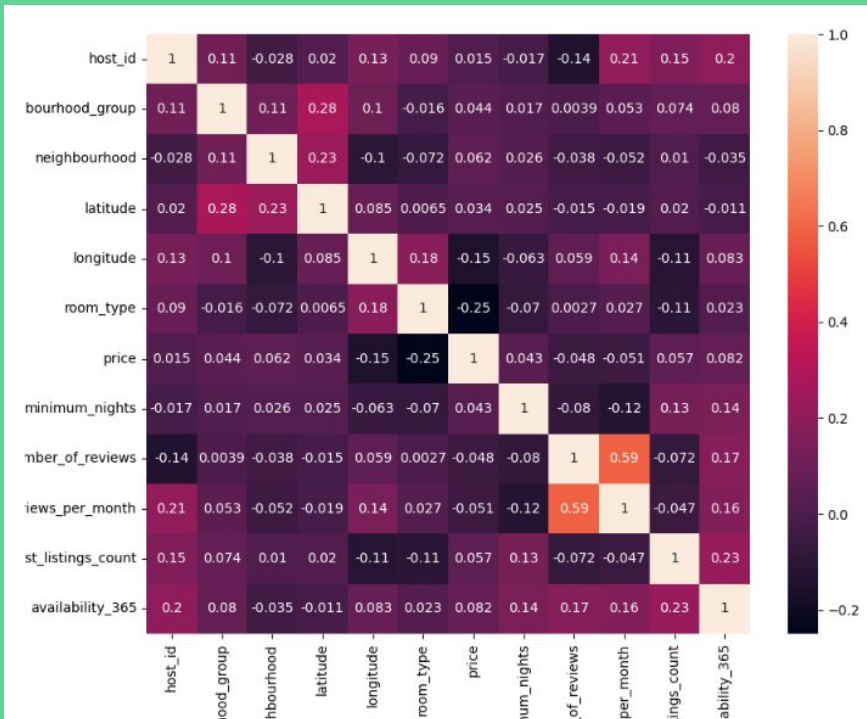
name	16
host_name	21
last_review	10052
reviews_per_month	10052

-Normalizaciones

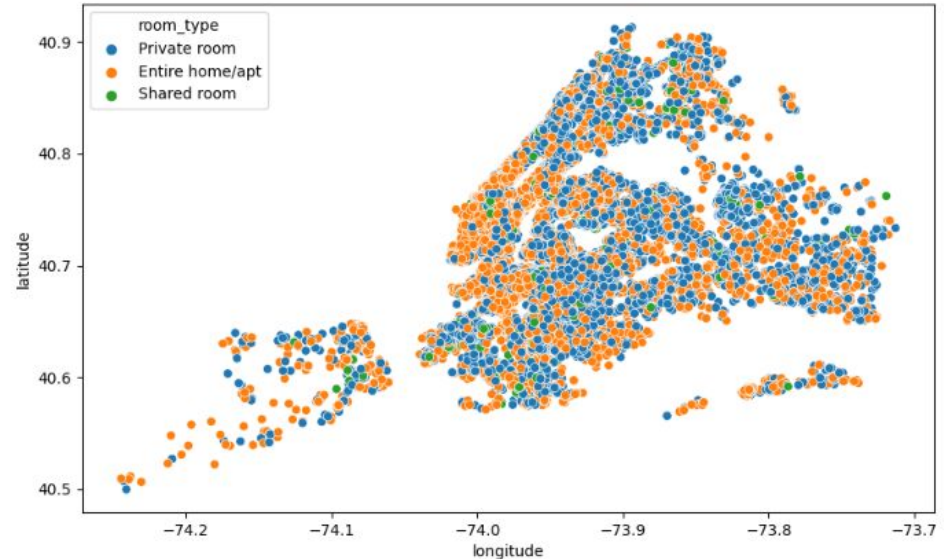
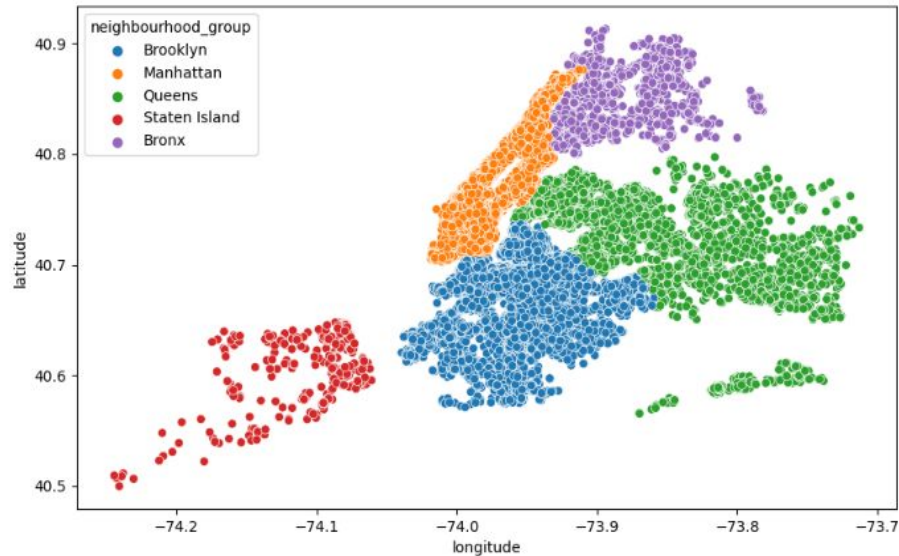
-Columnas eliminadas

-Tratado de outliers

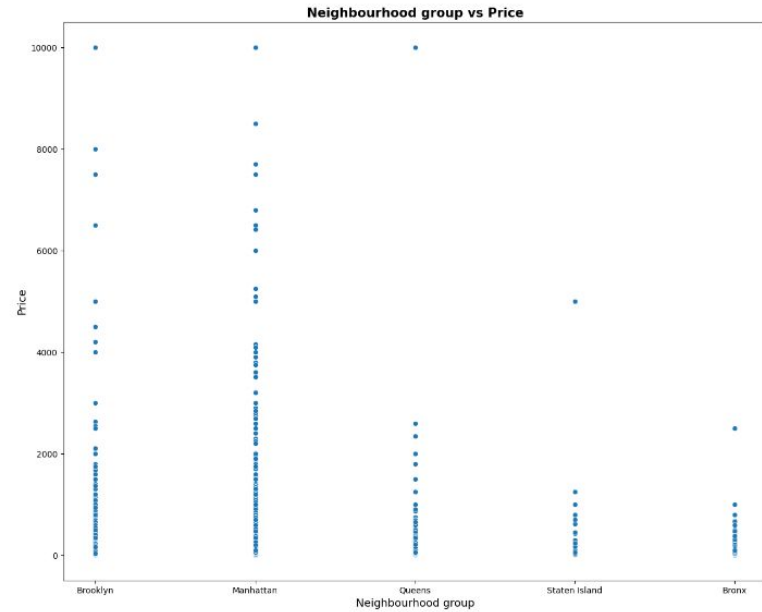
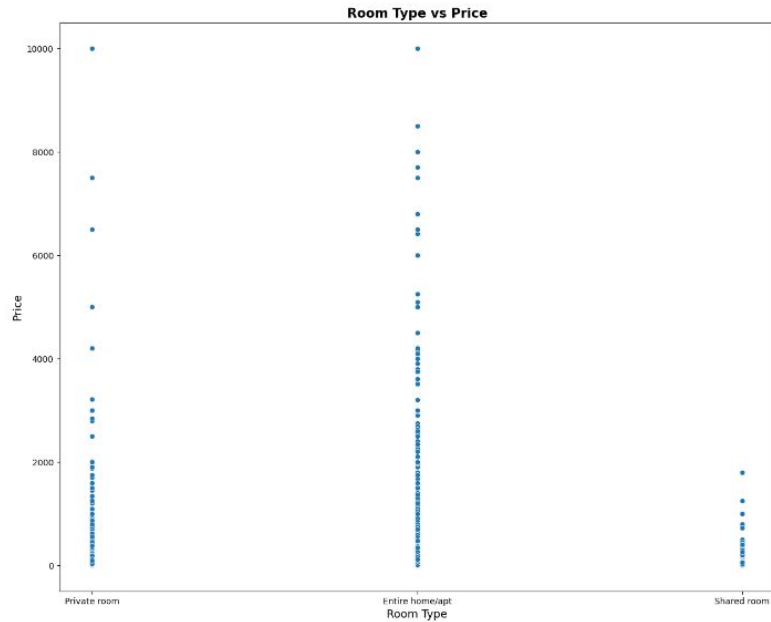
# Análisis de atributos



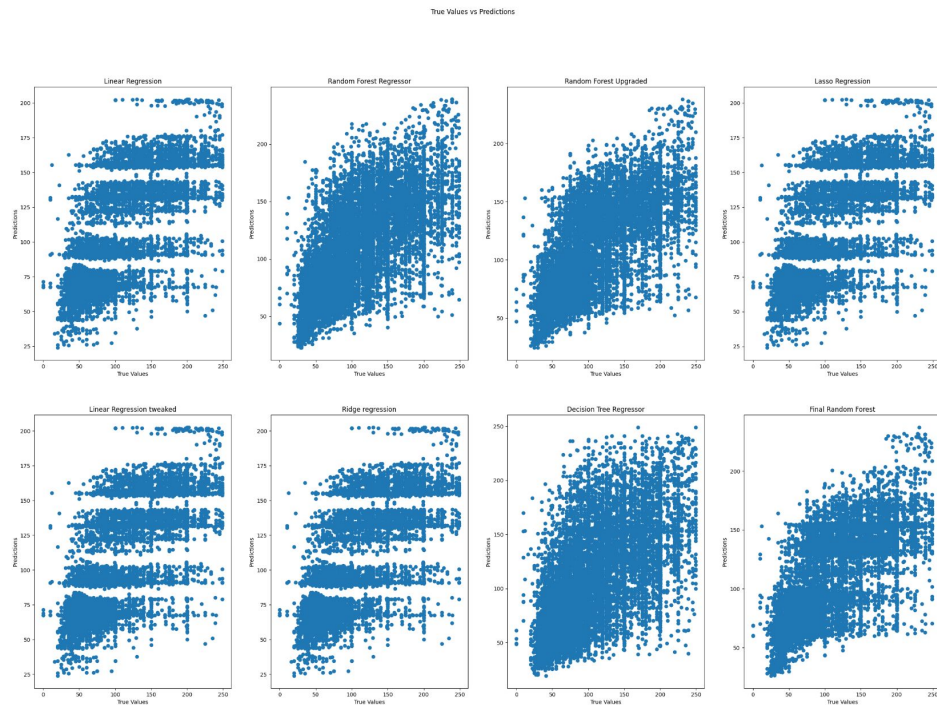
# Análisis de atributos



# Análisis de atributos



# Comparativa de modelos



Linear Regression  
R2: 0.485691  
MAE: 29.723851  
RMSE: 38.410790

Ridge Regression  
R2: 0.485691  
MAE: 29.723852  
RMSE: 38.410790

Decision Tree  
R2: 0.320126  
MAE: 33.376229  
RMSE: 44.162663

Lasso Regression  
R2: 0.485691  
MAE: 29.723873  
RMSE: 38.410791

Random Forest  
R2: 0.464046  
MAE: 29.804356  
RMSE: 39.210704

# Hiperparámetros

```
Gsr=RandomForestRegressor()  
param_grid = {  
    'n_estimators': [100, 150, 200],  
    'max_depth' : [10, 20, 30],  
    'min_samples_split' : [2,3,4,5]  
}  
  
Gsr = GridSearchCV(estimator=regRF, param_grid=param_grid, cv= 5)  
Gsr.fit(x_train, y_train)  
best_parameters = Gsr.best_params_  
best_score = Gsr.best_score_  
print(best_parameters)  
print(best_score)
```

```
{'max_depth': 10, 'min_samples_split': 4, 'n_estimators': 200}
```





Random Forest

R2: 0.516995

MAE: 28.499751

RMSE: 37.223466

# Presentación del estado de GitHub

 AuraKaz Update README.md	f38d12a 1 hour ago	🔒 3 commits
 .ipynb_checkpoints	Añadidos contenidos del repositorio	1 hour ago
 Dataset	Añadidos contenidos del repositorio	1 hour ago
 Images	Añadidos contenidos del repositorio	1 hour ago
 Analisis.ipynb	Añadidos contenidos del repositorio	1 hour ago
 README.md	Update README.md	1 hour ago

## Resumen

El dataset presenta 16 columnas con 6 de ellas siendo parámetros categóricos, otros son parámetros numéricos de los cuales se tuvo que aplicar una normalización a ciertos valores para computar una media y en otro caso se substituyeron los NaN por ceros por tal de rellenar los huecos debido a la lógica de los valores respecto a otra columna. Por otro lado tenemos que los valores categóricos se han convertido a numerales por tal de poder efectuar el análisis.

## Objetivos del dataset

Según se fue avanzando en el análisis de datos cada vez se concluyó mas que se efectuaría un intento de un modelo de predicción del precio en base al vecindario y el tipo de habitación, demostrandose generalmente bastante posible a lo largo del mismo.

## Experimentos

Durante este ejercicio se han efectuado múltiples experimentos, algunos se han desestimado según se avanzaban y algunos de ellos se han debido corregir debido a la inexperiencia que aún presento en este campo. Es por esto que entre los descartados encontramos el análisis de los precios y las reservas en base a las palabras seleccionadas en el nombre, experimento que se indicó como interesante debido a que podía indicar las tendencias de las personas a seleccionar X destinos en base a ciertas palabras. Otro experimento descartado porque se consideraba poco interesante fue el del impacto en las reservas que presenta la disponibilidad de los domicilios indicados. Finalmente el llevado a cabo y ya indicado es el de predecir precio en base a vecindario y tipo de habitación el cual comprende un análisis y prueba profunda de tanto el mejor modelo a nivel de resultados como de mejora mediante hiperparámetros de los mismos.

## Modelos

Durante el análisis se han usado múltiples modelos de regresión, los usados son los listados a continuación: -Linear Regression -Random Forest Regressor -Lasso Regression -Ridge Regression -Decision Tree Regressor

## Demo

Para obtener una demostración se recomienda ojear el fichero Analisis.ipynb, el cual gracias a github cuenta con los parámetros resultantes de la ejecución fijados con las gráficas obtenidas durante el mismo.

## Ideas para trabajar en un futuro

Por tal de obtener mas detalle sobre el valor de los datos se considera que el experimento de los nombres podría proporcionar información valiosa, también es posible que del mismo modo que se ha hecho un modelo predictivo fuese posible efectuar un modelo clasificador haciendo uso del precio y el tipo de habitación para determinar la población en la que se encuentra.

## Conclusión

El modelo que obtuvo el mejor resultado es el Random Forest, pienso que este es el que desde inicio iba a presentar mejores resultados debido a su capacidad de gestionar grandes volúmenes de datos. En comparación con los otros resultados, si es cierto que tenemos resultados relativamente parejos pero tenemos un problema mucho mas presente y visible de overfitting.



¡Muchas gracias por vuestra atención!

---