Precio de alquiler de alojamientos en AIRBNB - New York

Autor:	Lina Marcela Alvernia Arias
Comisión:	32695 Data Science
Institución:	CoderHouse

Contenido

1. Descripción del caso de negocio.
2. Objetivos.
3. Recolección de los datos
4. Preparación de los datos.
5. Análisis Exploratorio de Datos (EDA).
6. Elección del algoritmo.
7. Validación del modelo.
8. Conclusiones.
9. Bibliografía.

1. Descripción del caso de negocio

El progresivo y creciente aumento de la demanda de aplicaciones para la reserva de hospedaje se ha vuelto actualmente, un denominador común. Toda aquella persona que se encuentre interesada, ya sea en alquilar u ofrecer su propiedad para alquiler, coincide en la seguridad y facilidades que ofrecen este tipo de servicios.

Trabajamos para solucionar las inquietudes de un grupo inversor que encontró en el negocio de Airbnb un nicho de mercado. Este grupo, tiene como objetivo desarrollar su negocio en los Estados Unidos (USA) - New York, motivo por el cual quiere conocer cómo se desarrolla el mercado inmobiliario en esta zona, para así conocer como capitalizar mejor sus inversiones y poder **predecir el precio de alquiler** para los alojamientos que tengan disponibles, de acuerdo a sus características y oferta del sector.

Realizaremos un análisis exploratorio de datos para conocer cómo se distribuyen los precios promedios de alquiler en los distintos vecindarios de New York, buscando conocer cuáles son los vecindarios con mayor oferta de alojamientos, cuáles son los rangos de precios que se ofrecen y que relación tienen sus condiciones de servicio, la valoración de los usuarios, el año de construcción de la propiedad y si existe alguna incidencia en el costo del alquiler de dichas propiedades, para finalmente predecir los precios de aquellos alojamientos que se tengan disponibles para incluir en la plataforma.

2. Objetivos

- Analizar la distribución de los precios de alquileres según los distintos distritos de New York.
- Analizar tendencias de cada una de las variables para saber su influencia en el precio.
- Encontrar patrones entre las distintos características de los alojamientos en New York.
- Modelar un algoritmo de predicción de los precios de alquiler en New York.
- Brindar recomendaciones al grupo inversor.

Pasos en el procesamiento de la información.

01

Recolección de los datos, en este caso se extrae la información de Kaggle.

02

Preparación de los datos en Jupyter, Análisis Exploratorio de Datos (EDA), preprocesamiento, ETL, modelado de datos.

03

Elección de algoritmos y validación del modelo.

04

Presentación de la información, Deployment del modelo.

3. Recolección de los datos

- El dataset fue obtenido a través de Kaggle, en formato ".csv".
- Se cuenta con la información inicial de 102.599 alojamientos, registrados en la plataforma y 26 variables que los describen a cada uno de ellos.
- Variables principales sobre las que se centró el análisis: precio, tasa de servicio, año de construcción, noches mínimas, tipo de habitación, vecindario.
- Las variables del dataset originalmente eran del tipo "objeto", pero también se contaba con datos numéricos del tipo "int" o "float", los cuales fueron ajustados de acuerdo a lo requerido en el análisis.

N°	Campo	Descripción	Tipo
0	id	Identificador.	int64
1	NAME	Nombre del alojamiento.	object
2	host id	Identificador del host.	int64
3	host_identity_verified	Indica si el host tiene la identidad verificada.	object
4	host name	Nombre del Host.	object
5	neighbourhood group	Distritos de New York.	object
6	neighbourhood	Barrios.	object
7	lat	Latitud.	float64
8	long	Longitud.	float64
9	country	País.	object
10	country code	Código de país.	object
11	instant_bookable	Indica si el alojamiento permite hacer reserva inmediata.	object
12	cancellation_policy	Indica si el alojamiento cuenta con política de cancelación.	object
13	room type	Tipo de alojamiento.	object
14	Construction year	Año de construcción de la propiedad	float64
15	price	Precio de alojamiento por noche.	object
16	service fee	Precio de la tasa de servicio.	object
17	minimum nights	Indica el número mínimo de noches que debe ser alquilado.	float64
18	number of reviews	Número de reseñas del alojamiento	float64
19	last review	Última reseña	object
20	reviews per month	Reseñas por mes	float64
21	review rate number	Tasa de reseñas	float64
22	calculated host listings count	Recuento de listados de host calculado.	float64
23	availability 365	Disponibilidad para alquilar durante el año.	float64
24	house_rules	Reglas de la casa.	object
25	license	Licencia.	object

4. Preparación de los datos

- Para poder trabajar más fácilmente con el dataframe, se realizó el cambio de nombre de algunas columnas, reemplazando el espacio y agregando un guión bajo como separador, así mismo, dejando todo el campo en minúsculas, ejemplo: el campo *'Construction year'*, se reemplaza por: '*construction_year.'*

- Se eliminan las columnas*:* 'NAME', *'country', 'country code', 'id', host_name'* , 'calculated host listings count', 'availability 365', dado que el análisis se realizó para la ciudad de New York y sus distritos, por tal razón, estas columnas no son requeridas para dicho análisis, así mismo, se eliminaron las columnas:
  a) *License*: dado que tiene un total de 102.597 datos nulos, esto corresponde al 99,99% de los datos.
  b) *House rules*: dado que esta feature no sé tendrá en cuenta para la resolución de nuestro problema (tiene 52.131 datos nulos).
  c) *last_review:* dado que no es requerida para nuestro análisis.

- Se unificaron los valores de algunos distritos que se encuentran escritos de una manera diferente.
  a) *df = df.replace({"manhatan": 'Manhattan'})*
  b) *df = df.replace({"brookin": 'Brooklyn'})*
  c) *df = df.replace({"brookln": 'Brooklyn'})*

4. Preparación de los datos

Tratamiento de datos nulos:

- Se validó el total de datos nulos para cada una de las *features*, obteniendo el siguiente resultado:

```
df.isnull().sum()
```

id_host	0
host_identity_verified	289
neighbourhood_group	29
neighbourhood	16
lat	8
long	8
instant_bookable	105
cancellation_policy	76
room_type	0
construction_year	214
price	18150
service_fee	273
min_nights	409
number_reviews	183
reviews_month	15879
review_rate_number	326

- Se realizó el reemplazo de los NaN en las siguientes features de forma aleatoria.
  a) *price*
  b) *service_fee*
  c) *review_rate_number*
  d) *number_reviews*
  e) *lat*
  f) *long*

4. Preparación de los datos

Tratamiento de datos nulos:

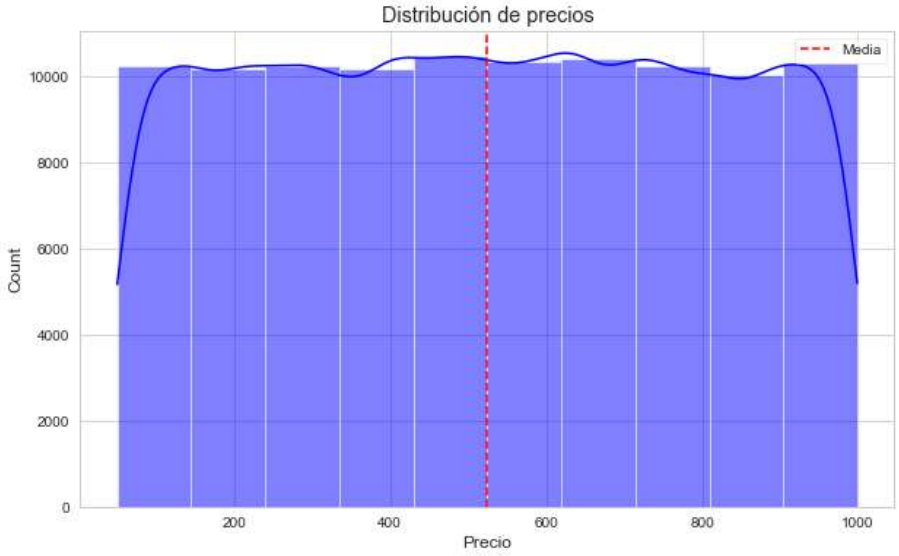
- Se realiza el siguiente tratamiento completando las instancias que tienen datos nulos:
  a) *neighbourhood:* se reemplazó los NaN por undefined, dado que se realizó el análisis por distrito (*neighbourhood_group).*
  b) *host_identity_verified:* se reemplazó NaN por unconfirmed.
  c) *instant_bookable:* se reemplazó NaN por False.
  d) *cancellation_policy:* se reemplazó NaN por strict.
  e) *min_nights:* se reemplazó NaN por el valor mínimo de 8 noches, hallado en el análisis descriptivo.
  f) *reviews_month:* se reemplazó NaN por 1.374, año promedio hallado en el análisis descriptivo.
  g) *construction_year:* se reemplazó NaN por 2012, año promedio hallado en el análisis descriptivo.

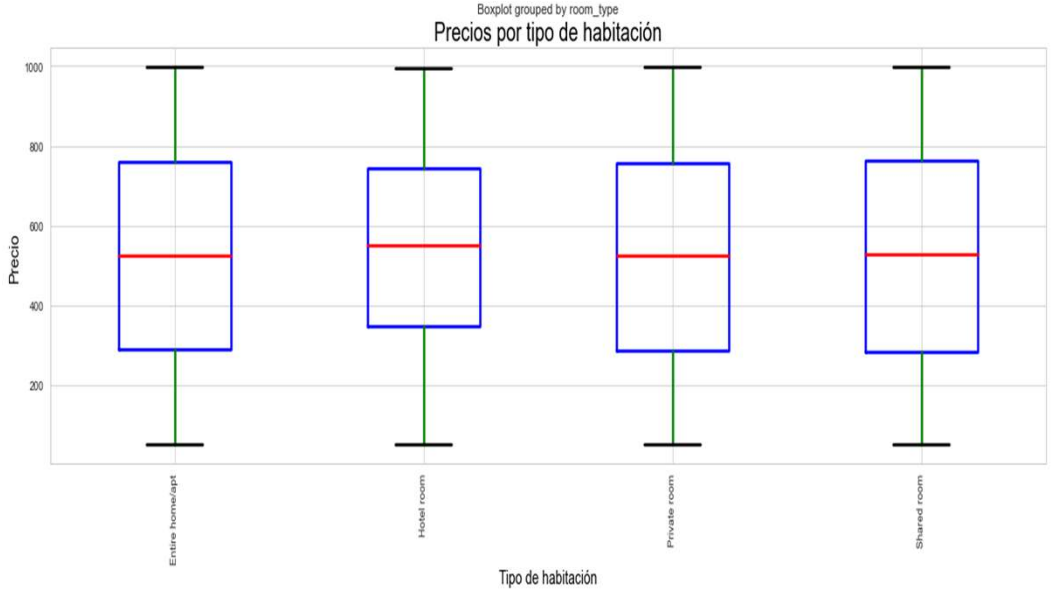
- Se realiza el reemplazo de los datos nulos de la variable *neighbourhood_group* obteniendo el valor a través de una tabla que contiene los distritos y los barrios.

- Finalmente, se valida que no se presenten datos nulos después de los cambios realizados.

5. Análisis Exploratorio de Datos (EDA)

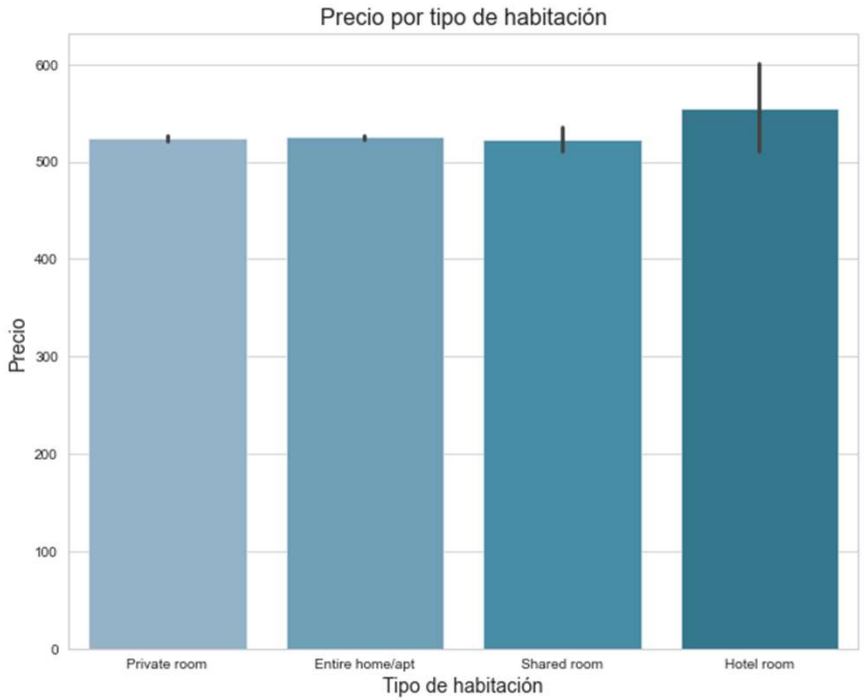
- La distribución de precios en New York es uniforme, por tanto, se pueden encontrar alojamientos con similar probabilidad entre los rangos de precios de 55 a 999 USD por noche, con una media de 524 USD como vimos en el análisis descriptivo de la variable.





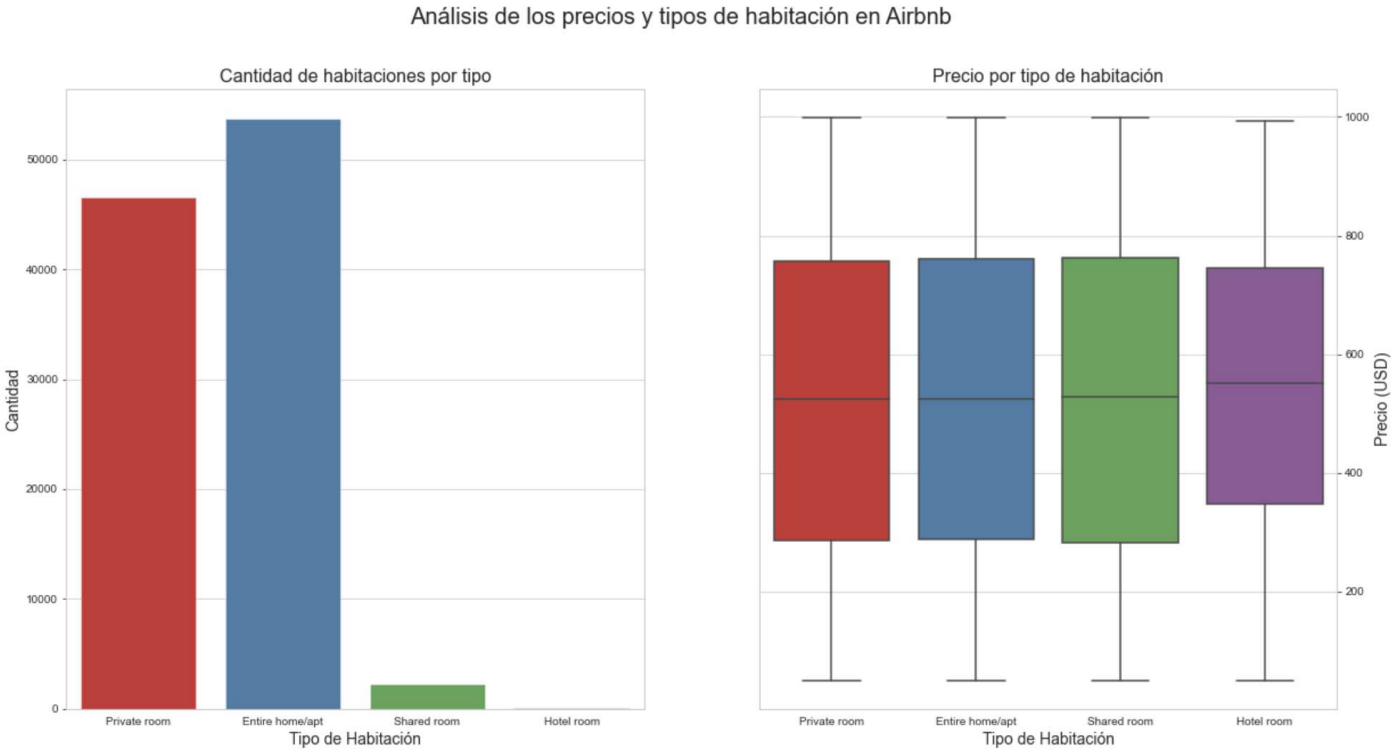
- De acuerdo a este boxplot, parece que no existe una mayor diferencia de precios entre los tipos de habitación.

5. Análisis Exploratorio de Datos (EDA)



- Graficamos un histograma para analizar en detalle la variación de precio por tipo de habitación, encontrando que los precios de habitaciones de hotel son un poco más altos en relación a los otros tipos de alojamiento (habitación privada, alojamiento entero, habitación compartida).

5. Análisis Exploratorio de Datos (EDA)

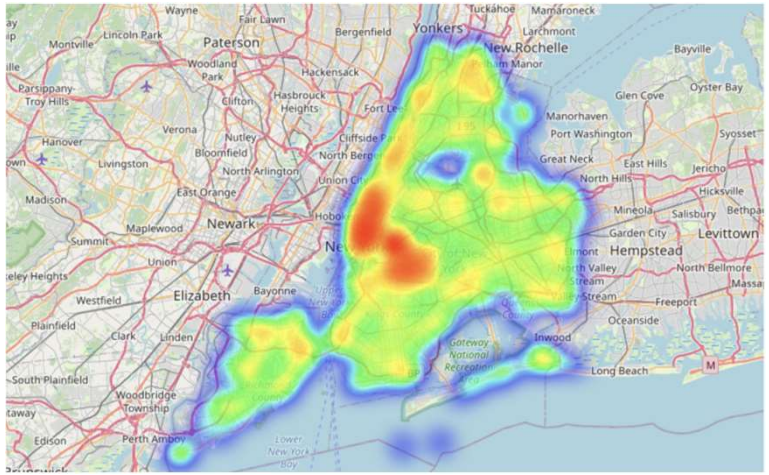


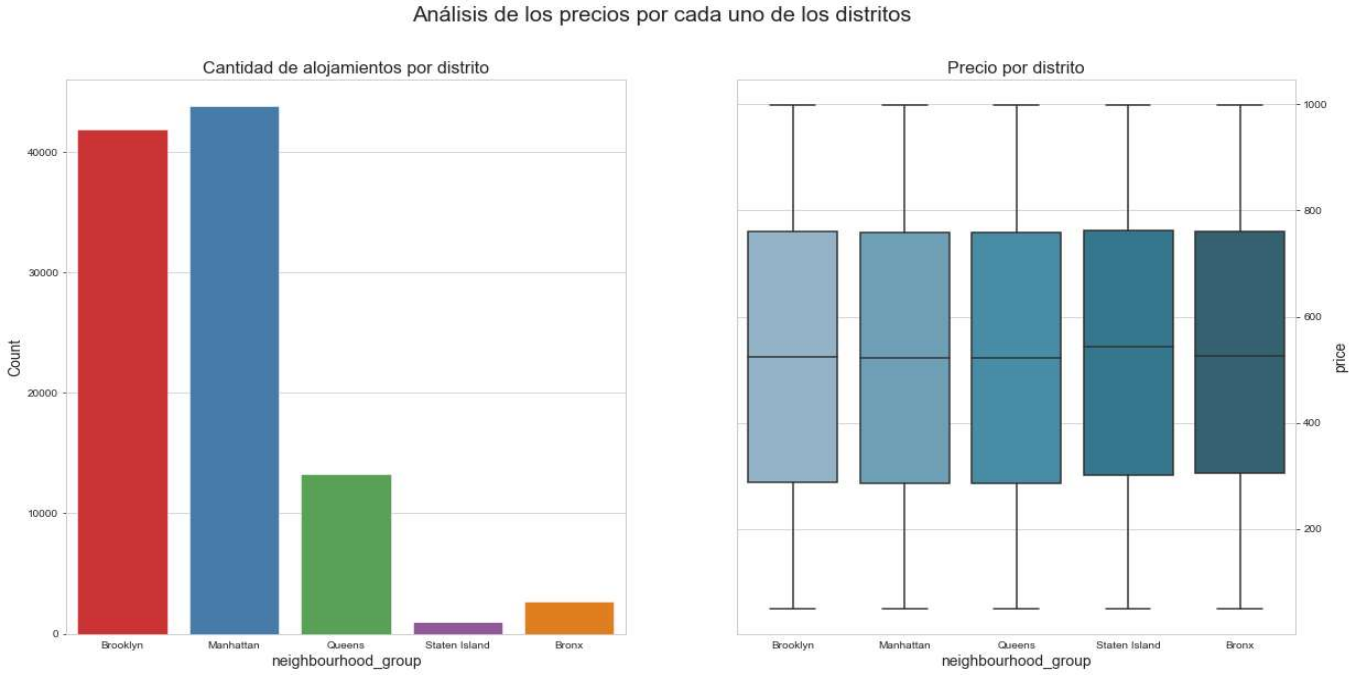
Así mismo, en este gráfico podemos observar, que las tendencias muestran una mayor oferta de aquellas propiedades que se alquilan de forma completa, seguidas por aquellas alternativas que ofrecen una habitación privada.

No representan una cantidad significativa las habitaciones compartidas ni las habitaciones de hotel.

5. Análisis Exploratorio de Datos (EDA)

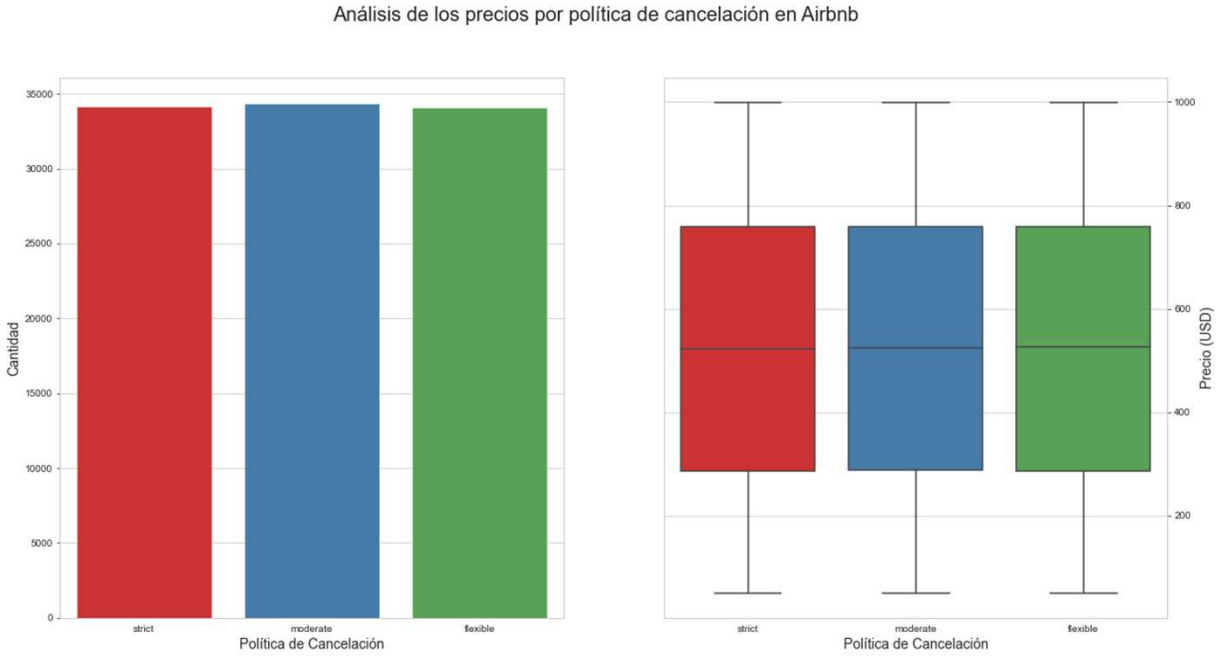
- Se grafican los alojamientos en el mapa de New York, confirmando que las zonas con mayor cantidad de alojamientos son Manhattan y Brooklyn. La menor cantidad de alojamientos se encuentran en Staten Island y Bronx.





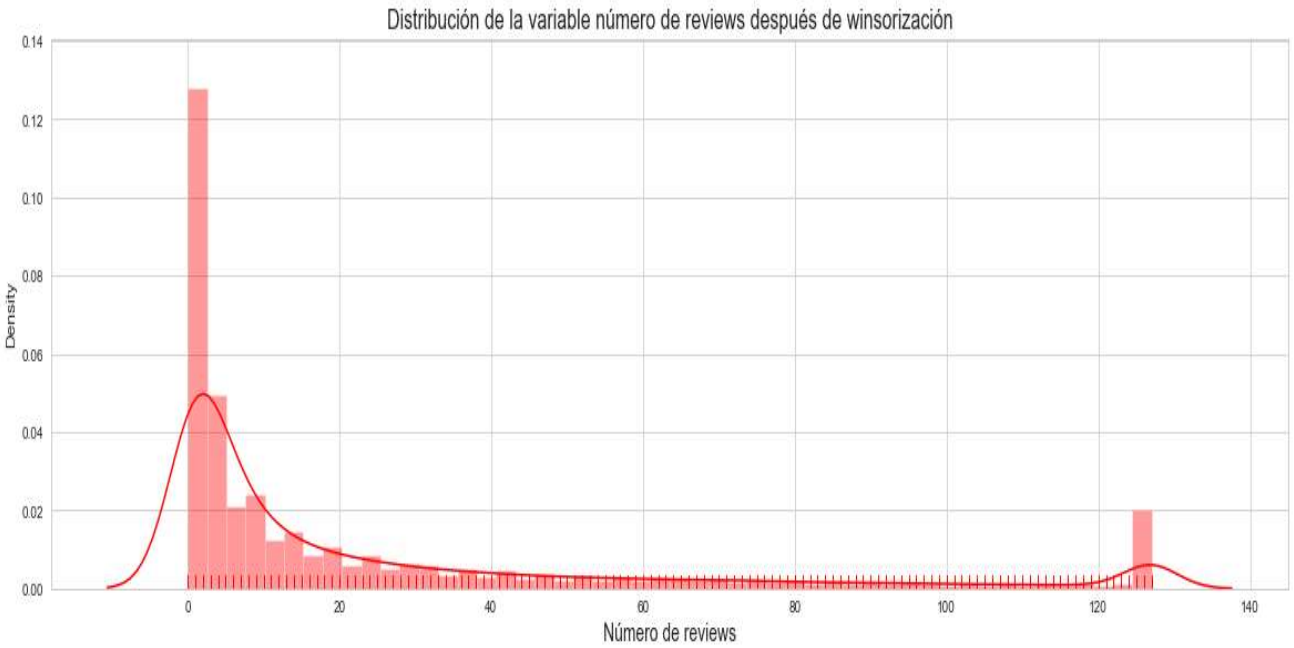
- Así mismo, se puede apreciar que no existe diferencia significativa en los precios de los alojamientos en cada uno de los distritos de New York.

5. Análisis Exploratorio de Datos (EDA)



- Con respecto a la política de cancelación, no se presentan diferencias entre aquellas que tienen políticas estrictas, moderadas y flexibles, ni en cantidad ni en precio.

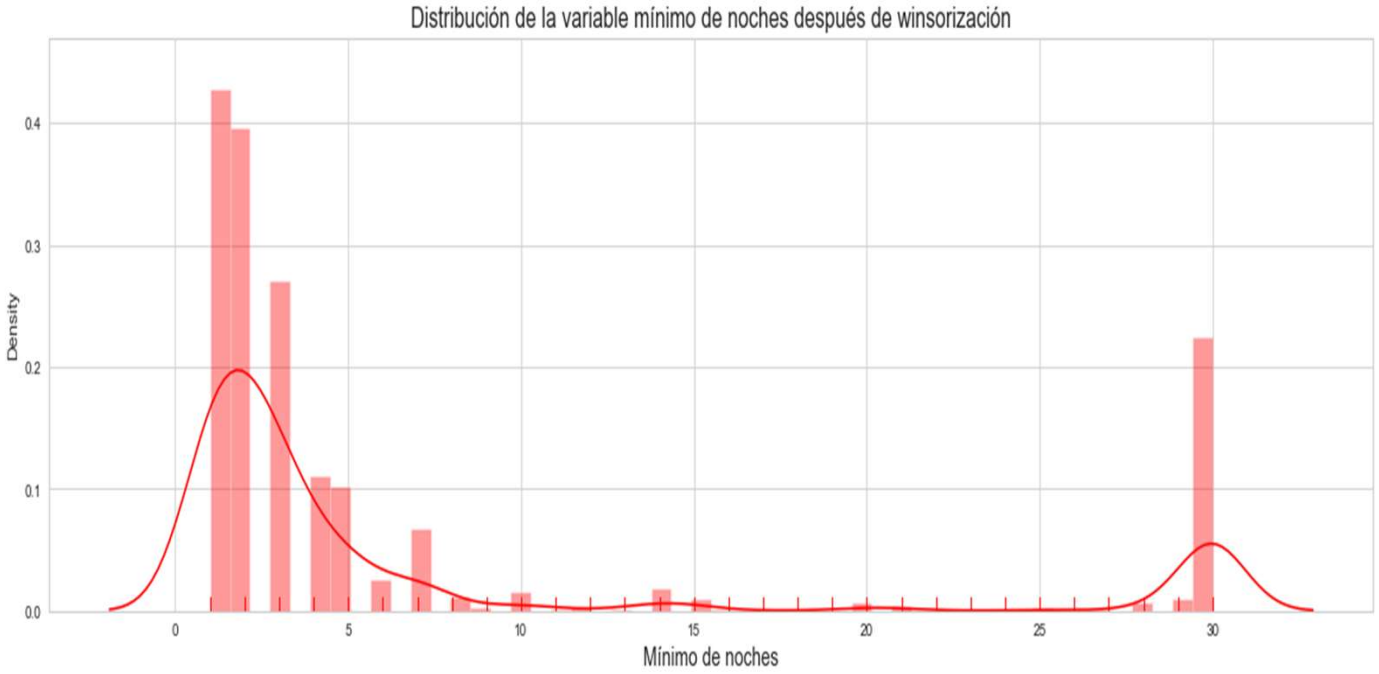
5. Análisis Exploratorio de Datos (EDA)



- Para observar el comportamiento de la variable número de reviews, se realizó un gráfico de distribución, encontrando que la gráfica tiene una inclinación hacia la izquierda, mostrando que existe un mayor número de alojamientos con pocas valoraciones, dado que la concentración en este punto es mayor.

- Se aplicó la técnica de winsorización para obtener una distribución más equilibrada de los datos, reduciendo la influencia de los valores atípicos en la variable. Sin embargo, la distribución sigue siendo inclinada hacia la izquierda debido al peso porcentual de esta concentración.

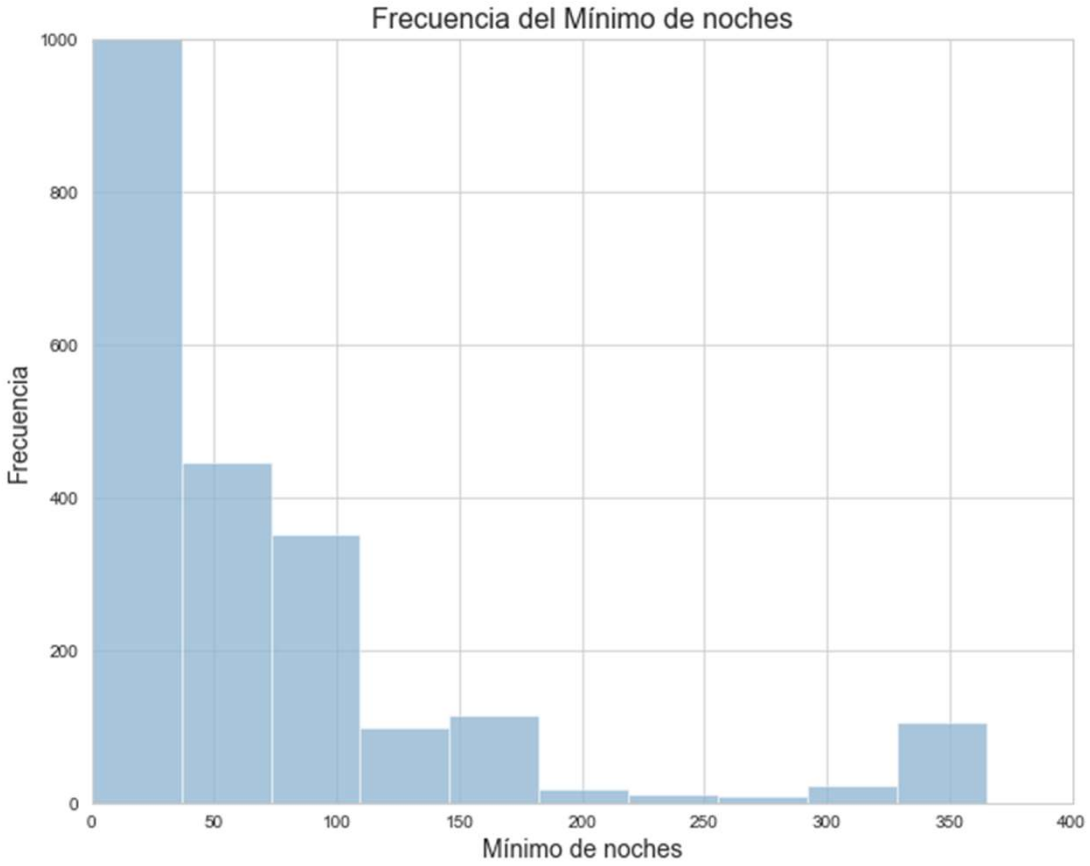
5. Análisis Exploratorio de Datos (EDA)



- Para observar el comportamiento de la variable mínimo de noches, se realizó un gráfico de distribución, encontrando que la gráfica tiene una inclinación hacia la izquierda, mostrando que existe un mayor número de alojamientos que requieren mínimas noches de estancia, sin embargo, existe otro grupo que requiere un alto número de estancia.

- Se aplicó la técnica de winsorización para obtener una distribución más equilibrada de los datos.

5. Análisis Exploratorio de Datos (EDA)



- Se realiza un histograma con la variable mínimo de noches, acotando el eje y de la frecuencia para poder observar el comportamiento de los demás grupos, por tanto, se observa que es mayor el número de alojamientos que requieren un mínimo de noches entre 0 a 200 noches.

- A partir de 200 noches, la oferta de alojamientos con este requisito disminuye considerablemente.

5. Análisis Exploratorio de Datos (EDA)

- Se analiza el precio por mínimo de noches y por grupo de vecindario y no se obtienen tendencias claras con estas variables.

- Sólo se puede deducir que existen alojamientos para todo tipo de exigencia en cuanto al mínimo de noches en cualquiera de los 5 distritos de la ciudad de New York. La oferta es uniforme.

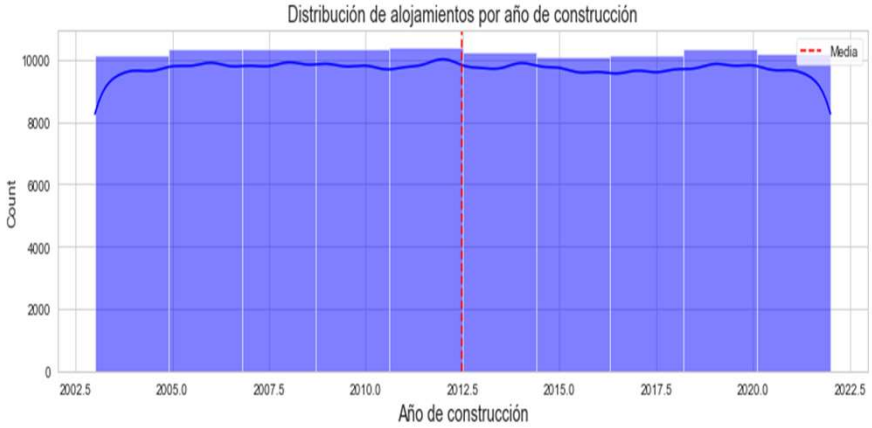
5. Análisis Exploratorio de Datos (EDA)

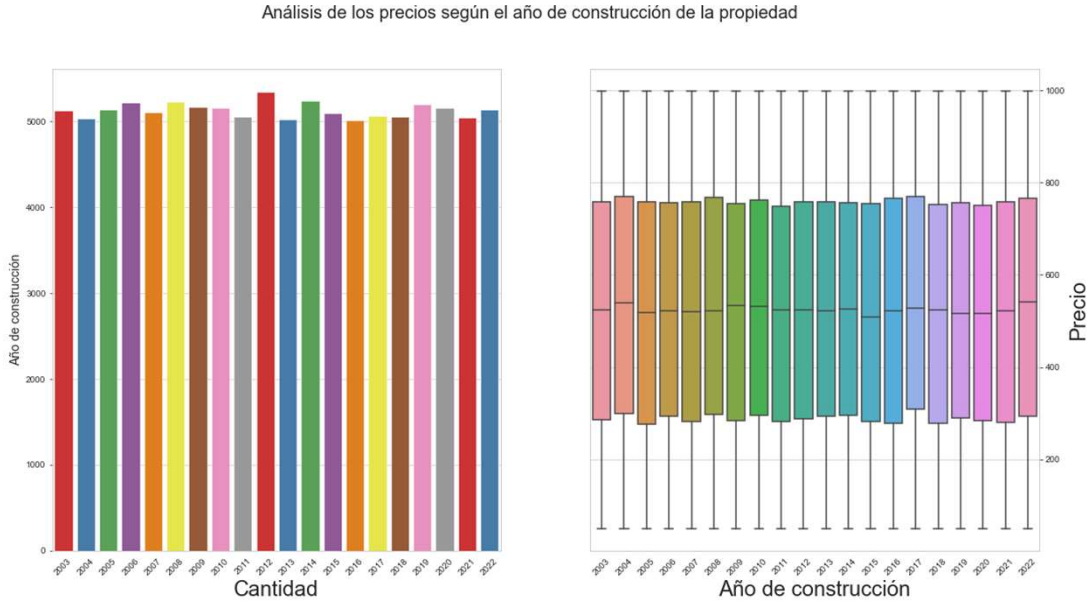
- Se procede a analizar las variables precio, número de reviews, política de cancelación, tipo de reserva y tipo de habitación para validar tendencias o grupos, sin embargo, se logra determinar que no se presenta una tendencia (ni positiva ni negativa), tampoco se observa una correlación entre sus variables.

- Se encuentran más alojamientos cuyas valoraciones oscilan en un rango de 0 a 200 en total, como se pudo observar en gráficos anteriores, encontrando que hay menor oferta de alojamientos con valoraciones mayores a este número.

5. Análisis Exploratorio de Datos (EDA)

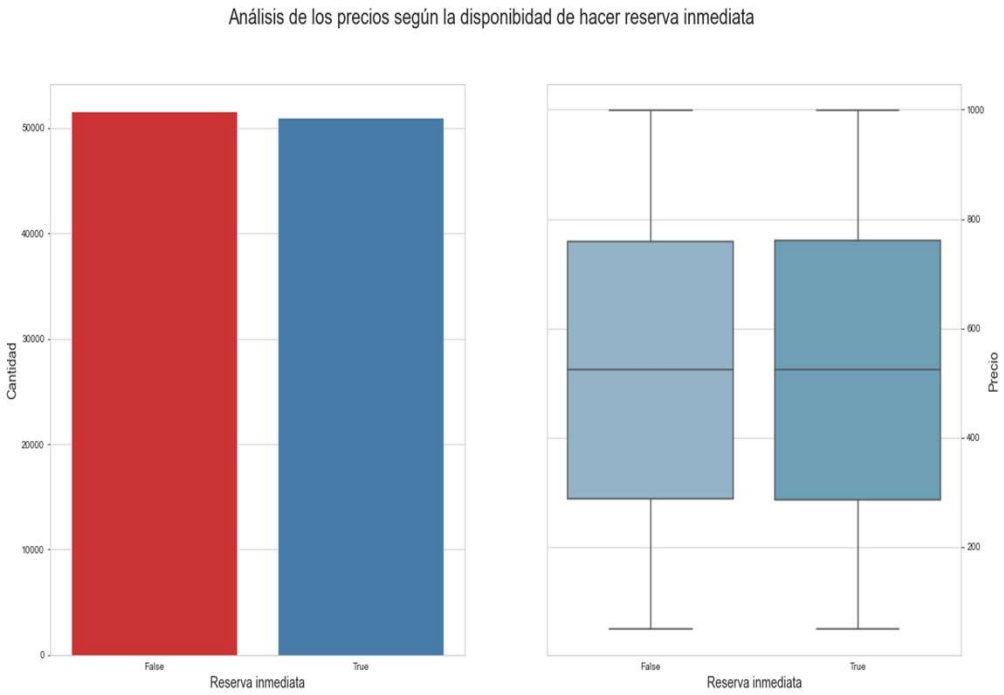
- La distribución de alojamientos por año de construcción en New York es uniforme, por tanto, se pueden encontrar alojamientos con similar probabilidad entre los rangos de años del 2002 al 2022, con una media de construcción del año 2012.

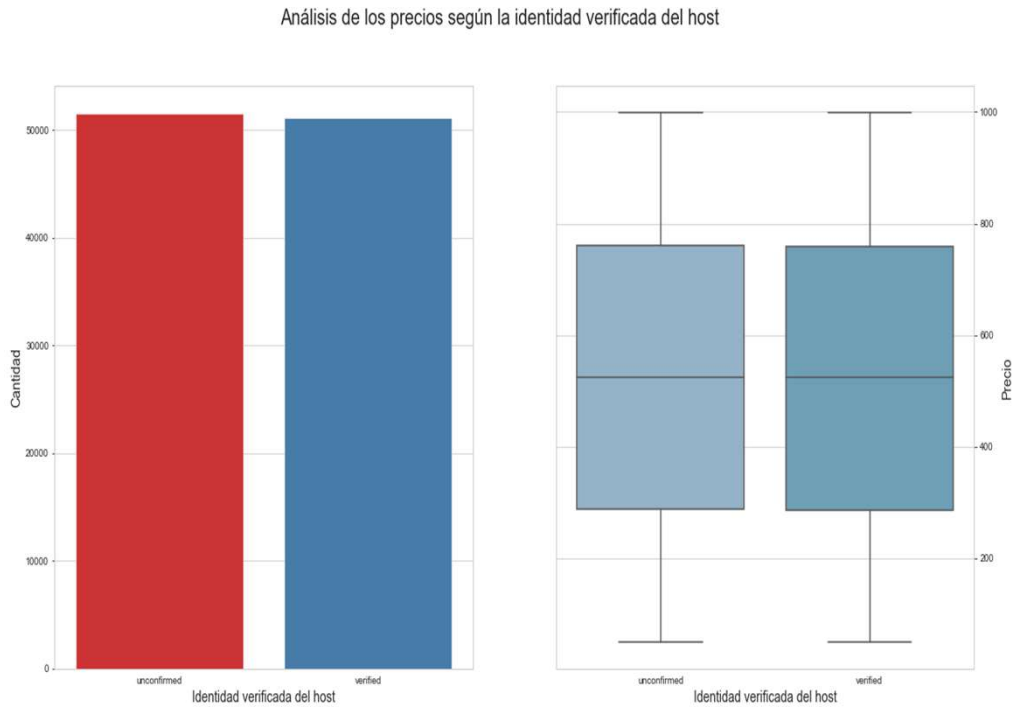




- De acuerdo a estos gráficos, no existe una mayor diferencia de precios entre los diversos años de construcción, su distribución es uniforme.

5. Análisis Exploratorio de Datos (EDA)





- De acuerdo a estos gráficos, no existen diferencias de precios cuando se analizan las variables: reserva inmediata y la identidad verificada del host.

6. Elección del algoritmo

- Para el desarrollo de este proyecto y con el fin de encontrar el mejor modelo para predecir el precio de los alojamientos y dar respuesta al grupo inversor, comparamos seis tipos de algoritmos:

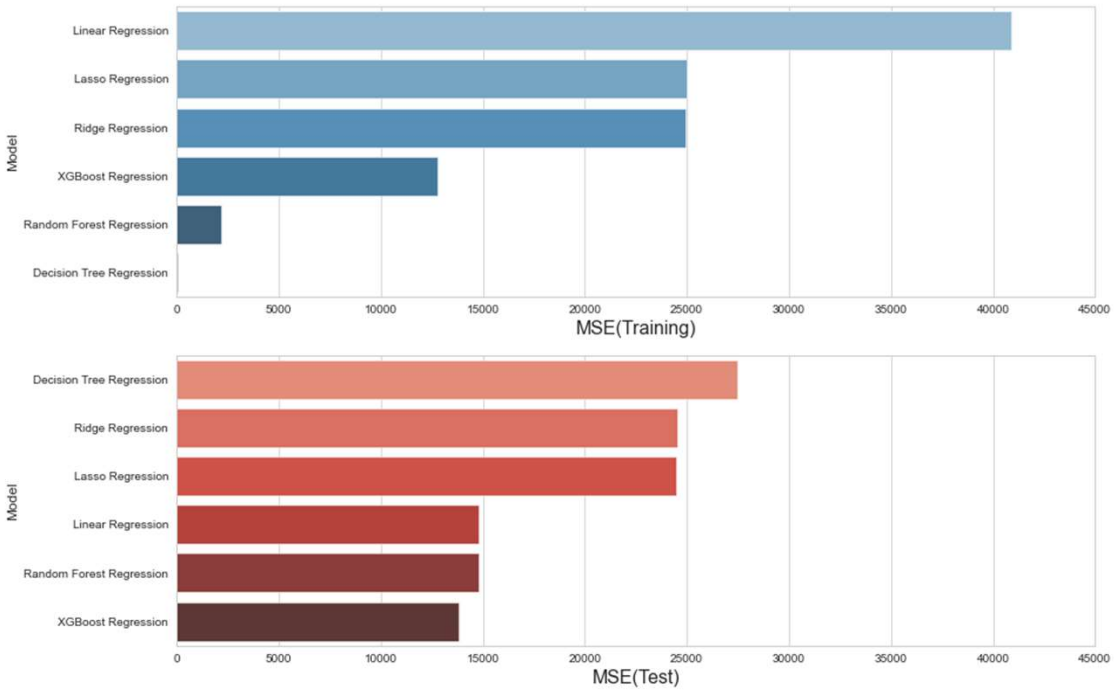
    1. Regresión Lineal.
    2. Random Forest.
    3. XGBoost.
    4. Árbol de Decisión.
    5. Ridge
    6. Lasso.

En los algoritmos Ridge y Lasso, se realizó el ajuste de hiperparámetros para modificar el rendimiento del modelo con el fin de lograr resultados óptimos en nuestra predicción, obteniendo de los rangos predeterminados los mejores parámetros para la optimización de estos modelos.

Finalmente, se realizó la validación del modelo a través del cálculo de 3 métricas como se aprecia a continuación:

7. Validación del modelo

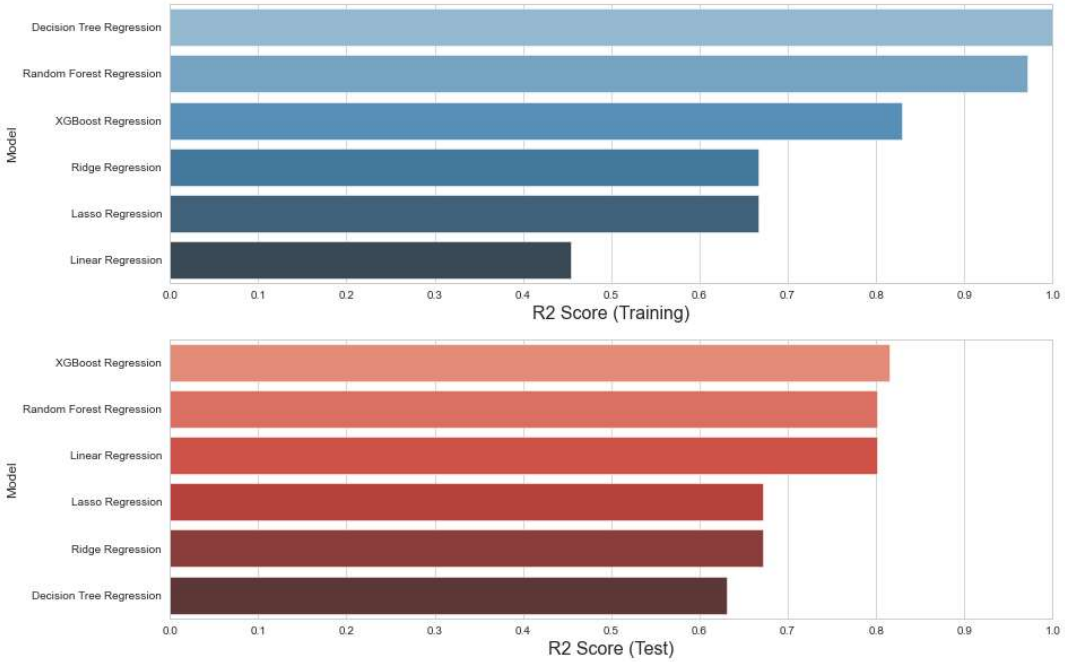
- Representación gráfica del rendimiento de los modelos:

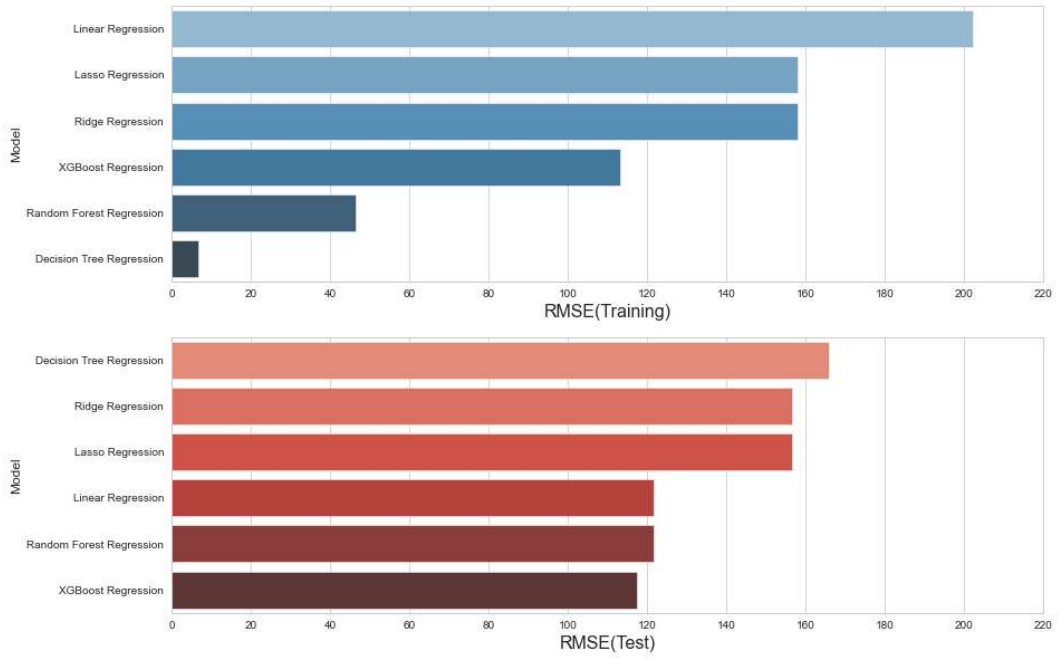


- Los datos que se presentan corresponden a las medidas de rendimiento (R2_Score, RMSE y MSE) obtenidas por diferentes modelos de regresión (*Linear Regression, Random Forest Regression, XGBoost Regression, Decision Tree Regression, Ridge Regression y Lasso Regression*) en dos conjuntos de datos: training y test.

7. Validación del modelo

- Representación gráfica del rendimiento de los modelos:





7. Validación del modelo

- De acuerdo al R2_Score, el modelo *Decision Tree Regression* tiene el mayor R2_Score en el conjunto de datos de entrenamiento, lo que sugiere que se ajusta muy bien a los datos, sin embargo, dado que este valor está muy cercano a 1, puede estar generando overfitting, y al validar esta misma métrica para el conjunto de datos de prueba, se tiene el menor valor de R2_Score confirmando este sobreajuste.

	Model	RMSE(training)	RMSE(test)	R2_Score(training)	R2_Score(test)	MSE(training)	MSE(test)
0	Linear Regression	202.203428	200.297686	0.454846	0.462323	40886.226166	40119.163062
1	Random Forest Regression	46.725771	122.062183	0.970889	0.800321	2183.297712	14899.176429
2	XGBoost Regression	113.045706	117.998203	0.829607	0.813396	12779.331673	13923.575980
3	Decision Tree Regression	8.402767	168.105079	0.999059	0.621268	70.606497	28259.317739
4	Ridge Regression	158.127448	156.548629	0.666607	0.671551	25004.289745	24507.473089
5	Lasso Regression	158.204032	156.509072	0.666284	0.671717	25028.515761	24495.089470

- El modelo *XGBoost Regression* tiene el mejor R2_Score en el conjunto de datos de prueba, lo que sugiere que realiza una buena predicción para datos no conocidos, y esta misma métrica para los datos de entrenamiento es similar, guardando las proporciones.

7. Validación del modelo

- Para el caso del RMSE (que representa la raíz cuadrada del error cuadrático medio y mide la cantidad promedio por la que los valores predichos difieren de los valores reales), podemos ver que el menor valor de RMSE lo tiene el modelo *Decision Tree Regression* para el conjunto de datos de entrenamiento, lo que sugiere un sobreajuste en estos datos.

- MSE (que es la media de los errores al cuadrado entre los valores predichos y los valores reales), muestra que el menor valor de MSE corresponde al modelo *Decision Tree Regression*, sin embargo, al validar esta misma métrica para los datos de prueba, tiene diferencias altas, por lo cuál sugiere un sobreajuste de los datos.

- Por otro lado, el modelo *XGBoost Regression* tiene el valor más bajo de MSE en el conjunto de datos de prueba, lo que sugiere que generaliza bien a datos no vistos, y este indicador también mantiene las proporciones con los datos de entrenamiento.

- De acuerdo a este análisis, el mejor modelo para predecir los precios de los alojamientos en New York es el XGBoost Regression , dado que cuenta con el mejor rendimiento en el conjunto de datos de prueba y mantiene la proporción con sus métricas de entrenamiento.

8. Conclusiones

En base al análisis de la información realizado, indicamos al grupo inversor lo siguiente:

- La distribución de los precios de alquileres según los distintos distritos de New York, es uniforme, no representa variación con respecto a las features de: distritos *(neighbourhood_grou*p), políticas de cancelación (*cancellation_policy*), número de reviews, tipo de reserva y año de construcción.

- Se presenta una pequeña variación en el precio de las habitaciones de hotel, dado que el límite inferior es más alto comparado con los alojamientos de tipo entero, habitación compartida y habitación privada, sin embargo, el valor promedio se mantiene uniforme.

- Las tendencias muestran una mayor oferta de aquellas propiedades que se alquilan de forma completa, seguidas por aquellas alternativas que ofrecen una habitación privada, sin embargo, se recomienda al grupo inversor la importancia de gestionar una base de datos de la demanda de alojamientos en New York para tomar la decisión del tipo de alojamiento a ofrecer.

8. Conclusiones

- Los distritos que tienen una mayor oferta de alojamientos son Brooklyn y Manhattan, sin embargo, se recomienda al grupo inversor gestionar la base de datos de la demanda de alojamientos de New York, para tomar la decisión del mejor distrito para ofrecer alojamientos.

- La variable que más influye en el precio es tipo de alojamiento, la cantidad mínima de noches promedio es 8 días, es decir por semanas.

- El mejor algoritmo para predecir el precio de alojamiento para las propiedades en New York es el *XGBoost Regression,* razón por la cuál este es el modelo recomendado para realizar el despliegue.

- Al haber una gran cantidad de oferta de alojamientos los precios promedios suelen ser uniformes en los diversos distritos de New York.

- Se recomienda al grupo inversor trabajar con la columna *House Rules*, para establecer una nueva feature correspondiente a aquellos alojamientos que aceptan mascotas en la propiedad y aquellos que no, para validar su oferta y precios.

9. Bibliografía

- https://pandas.pydata.org/docs/
- Manipulando datos perdidos en Python - Aprende IA
- https://matplotlib.org/
- https://scikit-learn.org/stable/
- https://www.cienciadedatos.net/machine-learning-python.html