Proyecto Modelos y Simulación de Sistemas I

Entrega 2

Laura Victoria Ramos Agudelo

Tutor

Raúl Ramos Pollán, Professor of Computer Science

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería de sistemas

Medellín

Planteamiento del Problema

La crisis de la vivienda holandesa es uno de los mayores problemas a los que se enfrentan los residentes. Debido a múltiples factores, como el crecimiento de la población y la escasez de trabajadores de la construcción, la disponibilidad de viviendas ha disminuido significativamente. Esta disminución ha llevado el alquiler a precios altísimos, lo que hace que muchos se pregunten si se están aprovechando de ellos.

Para responder a esta pregunta, el modelo debe predecir el alquiler de una casa a partir de sus características (es decir, ubicación, tamaño, instalaciones, etc.). El propósito de estos datos es poder investigar tendencias y patrones en el mercado de alquiler de bienes raíces en los Países Bajos. Con suerte, estos datos pueden explicar las situaciones actuales y ayudar a comprender lo que sucederá en este mercado.

Dataset o Base de Datos

El Dataset seleccionado es de una competición de Kaggle llamada Netherlands Accommodation Prices (FCG) y puede consultar en el siguiente enlace: https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/data. Esta base de datos contiene toda la información disponible en https://kamernet.nl/ para cada propiedad. El sitio web fue rastreado diariamente y si aparecía una nueva propiedad, se añadía a la base de datos. Si se encontraba una propiedad que ya existía, se añadía a las fechas de publicación.

Informe de Progreso y Planteamiento del Problema

He abordado diversas etapas en la investigación, desde análisis de los datos mediante creación de visualizaciones y estadísticas que arrojan luz sobre el mercado de alquiler de bienes raíces en los Países Bajos hasta un primer preprocesado de datos, añadiendo columnas clave que me servirán en el proceso de modelado. A continuación, resumo los hitos clave hasta el momento:

Análisis Exploratorio y Preprocesado

El análisis exploratorio de los datos revela información que el conjunto consta de un total de 27,915 filas y 34 columnas. Después de cambiar las columnas "firstSeenAt" y "lastSeenAt" a formato datetime tenemos que los tipos de nuestras columnas son: object, float64, int64 y datetime con 27, 3, 2 y 2 columnas respectivamente.

Debido a la limitada cantidad de información numérica disponible y la presencia de columnas con datos en formato de texto desafiante de procesar, como enlaces de imágenes o descripciones de publicaciones en línea, opté por crear columnas adicionales que enriquecieran los datos y

potencialmente influyeran en nuestra variable objetivo, el precio del alquiler, de la siguiente manera:

Precio de Alquiler por Metro Cuadrado (rent_per_areasqm): Se calculó el precio de alquiler por metro cuadrado dividiendo el precio de alquiler ("rent") por la superficie en metros cuadrados ("areaSqm"). Esta métrica ayuda a comprender la relación entre el precio de alquiler y el tamaño de la propiedad.

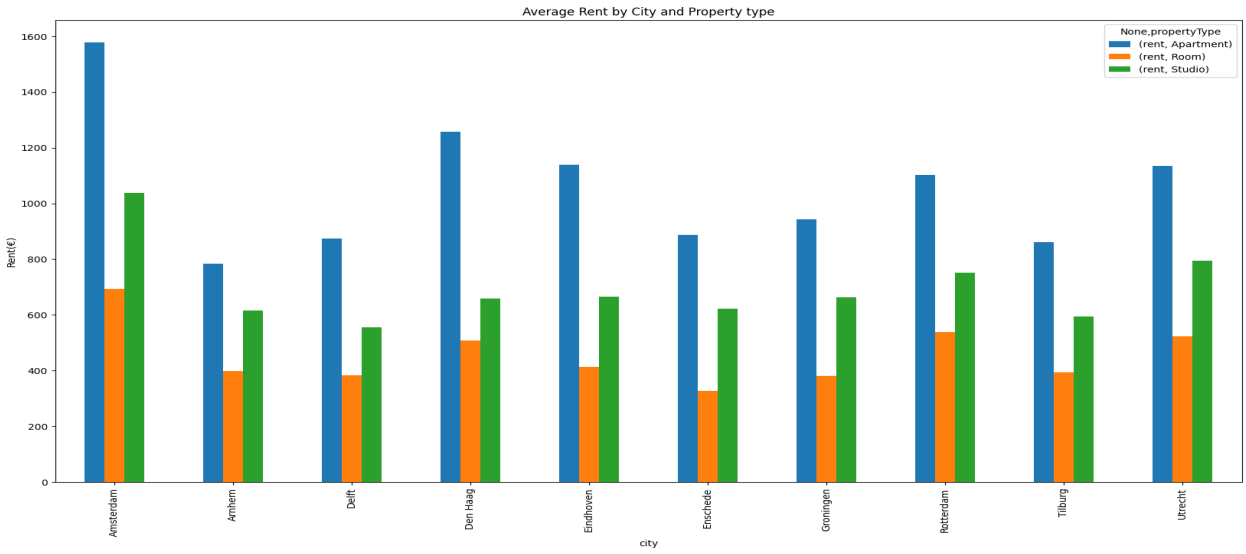
Distancia a las Ciudades Principales: Se calculó la distancia desde cada propiedad a las ciudades principales, lo que proporciona información valiosa sobre la ubicación de las propiedades en relación con los centros urbanos. Para este cálculo, se identificaron las 10 ciudades más frecuentemente listadas en el conjunto de datos y se definió un diccionario llamado "city_centers" que contenía las coordenadas geográficas (latitud y longitud) de la estación central de cada una de estas ciudades. Luego, se aplicó la fórmula de haversine. para calcular la distancia en kilómetros desde cada propiedad, hasta el centro de la ciudad correspondiente.

Por otro lado, se creó un nuevo DataFrame llamado rent_city_property_type que contiene estadísticas resumidas sobre el alquiler en las ciudades grandes, desglosadas por el tipo de propiedad.

city	propertyType	count	rent	rent_per_areasqm
Amsterdam	Apartment	1358	1578.620029	25.315516
	Room	3124	693.408131	50.505142
	Studio	371	1037.247978	37.343747
Arnhem	Apartment	262	784.083969	15.730438
	Room	528	398.659091	25.361136
	Studio	54	615.833333	22.678839
Delft	Apartment	27	874.037037	17.083956
	Room	655	383.911450	26.322397
	Studio	25	556.040000	20.164271
Den Haag	Apartment	389	1257.442159	18.461863
	Room	841	509.122473	33.403859
	Studio	84	659.273810	23.405312

El DataFrame tiene índices multinivel con las categorías de "city" y "propertyType", y columnas que muestran el conteo, el promedio del alquiler y el promedio del alquiler por metro cuadrado para cada categoría. Esto permite un análisis detallado de cómo varía el alquiler en función de la ciudad y el tipo de propiedad

Con dicha información se creó un gráfico de barras que muestra el promedio del alquiler en diferentes ciudades y tipos de propiedad. Los valores de alquiler se muestran en el eje vertical (Y), y las barras representan el promedio del alquiler para cada combinación de ciudad y tipo de propiedad. Esto proporciona una representación visual de cómo varía el alquiler en función de la ciudad y el tipo de propiedad.

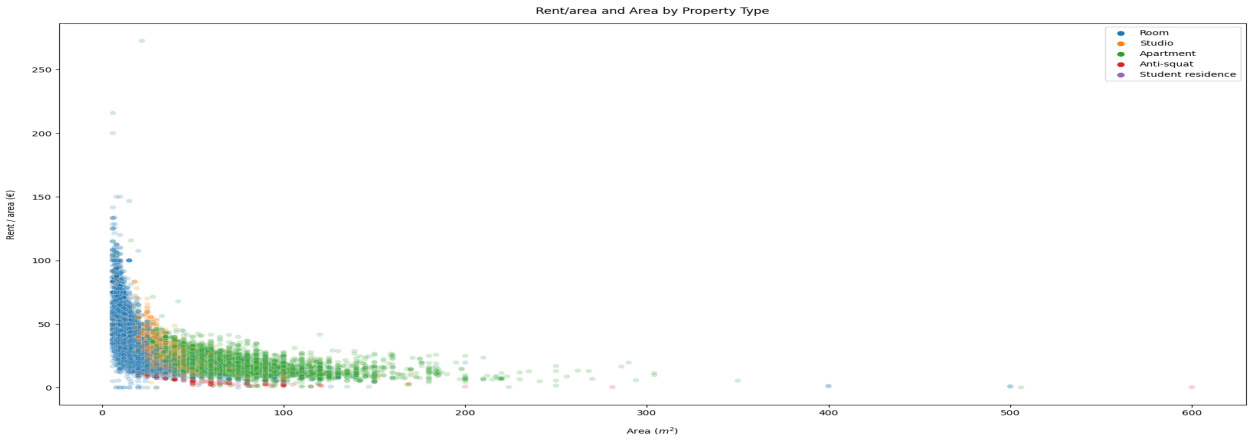


A partir de esta tabla y gráfico: si alguien desea vivir en un apartamento, Ámsterdam es la ciudad más cara con un promedio de 1600€ al mes, seguida de Den Haag con 1250€ al mes y Eindhoven, Rotterdam y Utrecht con alrededor de 1100€ al mes.

Para una habitación, el precio es de 700€ al mes en Ámsterdam, pero el precio por metro cuadrado será el más alto, a 50€ por metro cuadrado.

Para explorar la relación entre el tamaño, el precio y el tipo de propiedad, se utilizaron gráficos de dispersión que proporcionan una visualización efectiva de estos aspectos clave.

Relación entre Tamaño y Precio de Alquiler por Tipo de Propiedad



Las habitaciones, aunque pequeñas, son la opción más asequible. En contraste, los apartamentos varían en tamaño y precio, lo que sugiere una mayor diversidad de opciones. Es evidente que los apartamentos pueden presentar una mayor variabilidad en términos de precio y tamaño en comparación con las habitaciones.

Al observar los gráficos, es claro que a medida que el tamaño de la propiedad aumenta, también lo hace el precio. La relación entre el tamaño y el precio es directa. Sin embargo, existe una excepción en el caso de las propiedades "Anti-squat", que mantienen un precio constante independientemente del tamaño.

Al inspeccionar las variables categóricas, surgen varias preocupaciones sobre cómo procesar algunas de las columnas, como "matchAge," "matchLanguages," "coverImageUrl," y "descriptionNonTranslated." Estas columnas presentan una amplia variedad de categorías y valores únicos, lo que podría complicar su inclusión en un modelo analítico. En el caso de "matchAge," encontramos numerosas categorías que representan rangos de edades, pero también algunas categorías como "Not important - Not important." La diversidad de categorías podría dificultar la interpretación y el análisis. Con respecto a "matchLanguages," hay una gran cantidad de combinaciones de idiomas, lo que hace que esta columna sea compleja de manejar y puede requerir una codificación especial. Y lo mismo sucede con otras columnas, lo que plantea desafíos adicionales para su procesamiento en un modelo analítico.

Próximos Pasos y Retos Encontrados

Dada la complejidad y diversidad de estas columnas categóricas, es muy probable que no se tengan en cuenta al momento de comenzar a generar modelos predictivos, ya que podrían agregar complejidad innecesaria y sin aportar información significativa al análisis.

En resumen, se ha logrado un progreso significativo en el análisis de datos y exploración del mercado de alquiler de bienes raíces en los Países Bajos. Nuestro próximo paso será preprocesar columnas con valores nulos, y proceder con la construcción de modelos analíticos que permitan realizar predicciones precisas sobre los precios de alquiler, teniendo en cuenta todos los hallazgos y consideraciones realizadas hasta ahora y columnas clave que hasta el momento parecen ser: "areaSqm", "furnish", "propertyType", "internet", "kitchen", "living", "matchCapacity", "pets", "smokingInside", "dist_from_Amsterdam", "dist_from_Groningen", "dist_from_Rotterdam", "dist_from_Enschede", "dist_from_DenHaag", "dist_from_Utrecht", "dist_from_Eindhoven", "dist_from_Arnhem", "dist_from_Delft", "dist_from_Tilburg" y "rent_per_areasqm" .

Cita	Netherlands Accommodation Prices (FCG)
Referencia Estilo IEEE (2020)	[1] Kaggle. (s. f.). https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/overview