6.1 SOURCING OPEN DATA

DATA SOURCE: The dataset used for this assignment is an external data source available in an online platform called Kaggle. This dataset can be downloaded here.

DATA COLLECTION: The data was scraped from Immoscout24, the biggest real estate platform in Germany. Immoscout24 has listings for both rental properties and homes for sale, however, the data only contains offers for rental properties.

DATA CONTENTS: Data frame contains 268850 rows & 49 columns. The data set contains living area size, the rent, both base rent as well as total rent (if applicable), the location (street and house number, if available, ZIP code and state), type of energy etc. It also has two variables containing longer free text descriptions: description with a text describing the offer and facilities describing all available facilities, newest renovation etc.

DATA LIMITATION: There are lots of missing values and unnecessary information. There are typing mistakes too. There are duplicate columns too.

ETHICAL ISSUE: This data frame contains no personal information. So, no PLA security is required.

RELEVANCY & REASONING: This data frame contains all the necessary requirements for this project as it is open source, includes a geospatial component, meets the size and variable requirements.

Having resided in Germany for the past year and a half, during which I've encountered challenges in securing suitable housing, I've had a unique opportunity to delve into this comprehensive data framework. This experience has enabled me to extract valuable insights that shed light on various aspects of the rental market in the country.

DATA PROFILING:

VARIABLE	DESCRIPTION	TIME VARIABLE	QUALITATIVE/ QUANTITATIVE	DATA TYPE
State	State name	Invariant	Qualitative	Nominal
Service Charge	Auxiliary costs such as electricity or internet(euro)	Variant	Quantitative	Discrete
Heating type	Type of heating	Invariant	Qualitative	Nominal
Newly constructed	Is the building newly constructed?	Invariant	Qualitative	Binary
Balcony	Does the object have a balcony?	Invariant	Qualitative	Binary
Total rent	Total rent (usually a sum of base rent, service charge and heating cost)	Variant	Quantitative	Discrete
Year constructed	Year of construction	Invariant	Quantitative	Discrete

Has kitchen	Does the object have a kitchen?	Invariant	Qualitative	Binary
Celler	Does the object have a Celler?	Invariant	Qualitative	Binary
Base rent	Base rent without electricity & heating?	Variant	Quantitative	Discrete
Living space	Living space in sqm	Invariant	Quantitative	Discrete
Condition	Condition of the object	Variant	Qualitative	Nominal
Lift	Is elevator available?	Invariant	Qualitative	Binary
Type of flat	Type of flat	Invariant	Qualitative	Nominal
Number of rooms	Number of rooms	Invariant	Quantitative	Discrete
Floor	Which floor is the flat on?	Invariant	Quantitative	Discrete
Garden	Has a Garden?	Invariant	Qualitative	Binary
District	District name	Invariant	Qualitative	Nominal
City	City name	Invariant	Qualitative	Nominal
Date	Time of scraping	Variant	Qualitative	Ordinal

The following updates were made in the data frame:

Dropped columns:

- The following columns were dropped as it didn't contain useful information: 'telekomTvOffer', 'telekomHybridUploadSpeed', 'picturecount', 'pricetrend', 'telekomUploadSpeed', 'scoutId', 'noParkSpaces', 'firingTypes', 'geo_bln', 'yearConstructedRange', 'houseNumber', 'interiorQual', 'petsAllowed', 'street', 'streetPlain', 'baseRentRange', 'thermalChar', 'numberOfFloors', 'noRoomsRange', 'livingSpaceRange', 'geo_krs', 'description', 'facilities', 'heatingCosts', 'energyEfficiencyClass', 'lastRefurbish', 'electricityBasePrice', 'electricityKwhPrice', 'geo_plz'

Renamed columns:

- 'regio1' was changed to 'state'.
- 'regio2' was changed to 'district'
- 'regio3' was changed to 'city'
- 'noRooms' was changed to 'numberOfRooms'
- 'newlyConst' was changed to 'newlyConstructed'

Changed data types:

- ‘numberOfRooms’ changed from a float to an integer data type as number of rooms cannot be float.
- ‘yearConstructed’ changed from float to integer data type as year of construction cannot be float.
- ‘floor’ changed from float to integer data type as floor cannot be float.

Mixed type data:

- ‘heatingType’, ‘condition’, ‘typeOfFlat’ datatype were changed to string.

Missing values:

- Replacing missing values from ‘serviceCharge’ from its mean value i.e. 151
- Replacing missing values from ‘totalRent’ from its mean value i.e. 902
- Replacing missing values from ‘yearConstructed’ from its median value i.e. 1973
- replacing missing values from ‘floor’ from its median value i.e. 2

Key questions:

1. What is the current status of rental market in Germany: Is it stable, declining or rising?

2. What factors influence the rental prices the most (location, size, amenities)?

3. Which cities or regions in Germany have highest and lowest average rentals?

4. Is there any correlation between rental price and factors like location, size and amenities?

5. What type of rental properties are most common in Germany: house, apartment, studio etc.? Does different property type have different price ranges?

6. Based on historical data, can we forecast upcoming rental trends??

7. Are there noticeable seasonal trends in the rental market? Do price trend to fluctuate throughout the year?