6.1. Project information, Sandra Wienecke

Data Source

Name of Data sets

- https://www.kaggle.com/corrieaar/apartment-rental-offers-in-germany

Summary

- **Data Source:** The data was scraped from a private person, 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. At a given time, all available offers were scraped from the site and saved. This process was repeated three times, so the data set contains offers from the dates 2018-09-22, 2019-05-10 and 2019-10-08.
- **Data Collection:** Because of scraping, the data belongs to www.immobilienscount24.de and is for research purposes only. The data was created with R. The scraping process is described: https://www.samples-of-thoughts.com/2018/scraping-the-web-or-how-to-find-a-flat/
- **Data Contents:** This data contains information about apartment rental offers in Germany, including their characteristics, e.g. location, costs, building fabric & energy certificate, equipment.

Why did I choose this dataset?

I live in Berlin. Berlin is a big city that used to have the reputation of having cheap rents. While other cities are certainly still much more expensive, rents in Berlin have risen significantly. I am interested in what rents are like in German cities.

Data Profile

Data cleaning & consistency checks

- The original data set contains 49 columns and 268850 rows.
- Dropping columns:
    - I dropped some columns which are double but with different names: geo_bln geo_krs, streetPlain
    - I dropped both columns which have detailed adress information which are not needed here: houseNumber, street
    - I dropped columns with too many value for a categorical column/ : description, facilities

    - I dropped columns where more than 25% of the values are missing:

      'telekomHybridUploadSpeed', 'noParkSpaces', 'condition', 'interiorQual','petsAllowed', 'thermalChar', 'numberOfFloors', 'heatingCosts', 'energyEfficiencyClass', 'lastRefurbish', 'electricityBasePrice', 'electricityKwhPrice'

- Renaming columns: I would like to have renamed all the columns for them to make it more intuitive. So it's just not worth to rename all the columns. So I only renamed ones that I thought might be used most important. I renamed :
    - 'regio1' to 'state'
    - 'geo_plz' to 'postcode'
    - 'regio2' to 'city/county'
    - 'regio3' to 'city district'

- Fixing mixed-type columns: the following columns have mixed-type values. I changed all the type of them to string.
    - heatingType
    - telekomTvOffer
    - firingTypes
    - condition
    - interiorQual
    - petsAllowed
    - typeOfFlat
    - energyEfficiencyClass

- Missing values:Because of dropping columns with more than 25%, there are no missing values left.
- Duplicate values: There were no full duplicate values.

Basic descriptive statistics

Rows: 268.850
Columns: 30
Total record count: 8.065.500

Table

	serviceCharge	picturecount	pricetrend	telekomUploadSpeed	totalRent	yearConstructed
count	268850.000000	268850.000000	268850.000000	268850.000000	2.688500e+05	268850.000000
mean	151.206113	9.791958	3.389001	28.804928	9.013315e+02	1966.400590
std	304.308647	6.408399	1.958168	15.290063	3.063151e+04	41.709858
min	0.000000	0.000000	-12.330000	1.000000	0.000000e+00	1000.000000
25%	96.000000	6.000000	2.000000	10.000000	4.950000e+02	1958.000000
50%	140.000000	9.000000	3.389001	40.000000	7.410000e+02	1966.400590
75%	185.000000	13.000000	4.560000	40.000000	9.013315e+02	1992.000000
max	146118.000000	121.000000	14.920000	100.000000	1.575154e+07	2090.000000

Out[81]:

scoutId	yearConstructedRange	baseRent	livingSpace	baseRentRange	postcode	noRooms	floor	noRoomsRange	livingSpaceRange
8500e+05	268850.000000	2.688500e+05	268850.000000	268850.000000	268850.000000	268850.000000	268850.000000	268850.000000	268850.000000
9697e+08	3.714544	6.941294e+02	74.355548	3.765256	37283.022235	2.641261	2.122405	2.571542	3.070790
0093e+07	2.430343	1.953602e+04	254.759208	2.214357	27798.037296	2.633440	3.269730	0.937594	1.407127
7174e+07	1.000000	0.000000e+00	0.000000	1.000000	852.000000	1.000000	-1.000000	1.000000	1.000000
6910e+08	2.000000	3.380000e+02	54.000000	2.000000	9128.000000	2.000000	1.000000	2.000000	2.000000
1584e+08	3.714544	4.900000e+02	67.320000	3.000000	38667.000000	3.000000	2.000000	3.000000	3.000000
7688e+08	5.000000	7.990000e+02	87.000000	5.000000	57072.000000	3.000000	3.000000	3.000000	4.000000
7117e+08	9.000000	9.999999e+06	111111.000000	9.000000	99998.000000	999.990000	999.000000	5.000000	7.000000

Freq. Table: See Jupyter notebook

Limitations and ethical considerations

The data from the dataset is from a short period of time as the scrap via Immoscout24 was only done three times for the dataset (May 2019, Oct 2019 and Feb 2020). For a more comprehensive picture, data from a longer period would be more helpful.

Furthermore, the data set only contains data from the Immoscout24 platform. In Germany, however, there are of course more platforms on which flats are offered. For example Immowelt, Ebay Kleinanzeigen, Immonet, Wohnungsbörse24. For a more comprehensive picture, data from different platforms would be interesting.

It is not possible to tell from the data who use the Immoscout24 platform. Do the users represent a good profile of society, so that the housing offers are also reflected in this? Users are understood here to be those looking for housing and those offering housing. And the question arises as to how this applies to other housing platforms.

The data on Immoscout24 is entered manually by people who offer flats. Therefore, it can happen that some information is forgotten when the offer is made and is only noticed or discussed later during the viewing of the flat

Define Questions

Clarifying questions:

- Which city have most expensive flats?
- Where can people live most cheaply? Where are the 10 affordable flats?
- Which characteristics do expensive and cheap flat have?
- Where are the most and the fewest flats offer?

Funneling questions:

- Are flats with almost the same characteristics in the same price level?
- Does the city district have an influence on the price? baseRent city district
- Are renovated flats in the same price range as first time use flats? Does the year of construction have an influence on the price? yearConstructed, baseRent
- Are bigger flats more expensive than smaller flats? Number of rooms, livingSpace, baseRent
- Do the number of plat offers in a region/city have an influence on the price? City & Scoutid, baseRent