**Data Sources**

Immoscout24 dataset

**About:** I chose this data because I’m interested in the cost of housing, and I wanted to work with a large dataset that would require analysis in Python and focus on a topic of interest in Europe, where I’ll be focusing my job search.

**Source:** This is external data from Immoscout24, the largest real estate website in Germany.

**Collection method:** A private user scraped the data from Immoscout24’s website on four dates between 2018 and 2020 and [posted it to Kaggle](https://www.kaggle.com/datasets/corrieaar/apartment-rental-offers-in-germany).

**Contents:** The dataset contains a range of information about all rental units live on Immoscout24 at the time of scraping, including total rent, location, unit size, apartment size, building age, type of heating, floor number and other variables listed in full below.

**Limitations:** The raw data includes a number of errors that likely resulted from scraping and processing and/or human error in the listings. I cleaned it thoroughly to address as many incorrect or suspicious values as I could, focusing mostly on outliers in numerical variables, but errors likely remain. The rental data is several years old and doesn’t reflect impacts of the COVID-19 pandemic, recent inflation or other events that have impacted rental pricing. Finally, the dataset only represents rental units that were posted on Immoscout24 during the four scraping dates, so it doesn’t encompass occupied rental units or available units that weren’t advertised on the site.

**Ethics:** The data contains no personal information, but it belongs to Immoscout24 and may only be used for research purposes.

**Relevance:** This data will allow me to investigate the cost of rent in Germany depending on location, apartment size and other factors that could prove interesting. Since the data was collected over the course of several years, I can also gauge whether average rent changed over time and at what rate.

**Statistisches Bundesamt dataset**

**About: I created this dataset myself by combining and transforming two datasets from the German statistical office. I’m interested in population data and wanted to incorporate it in this project, so I decided to examine the potential relationship between rental costs and total population, population growth and housing units relative to population.**

**Source:** This is external data from **Genesis Online, the online data portal for the German federal statistical agency Statistisches Bundesamt.**

**Collection method**: Statistisches Bundesamt conducted a formal census via statistical survey in 2011 and 2022. (The 2022 data isn’t available yet.) The other census numbers come primarily from population registration – German residents are required to register with local authorities when they move. The government collects data on residential units via mail on a regular basis and updates it annually by statistical model.

**Contents:** The dataset contains annual estimates of population, number of apartments and total living space (in 1,000 square meters) for all 16 German states and Germany as a whole between 2000 and 2021. I derived two additional columns representing residents per apartment and living space per resident for each state annually.

**Limitations:** Though population registration is required by law, some residents don’t do it, and there’s a lag in the data for people who’ve moved to another state or entered Germany and haven’t registered yet. The formal census process also misses some people, particularly in Germany, where the public is generally distrusting of the census process. The method of collection for residential building data is also subject to reporting error and lack of responses, and the annual calculations of residential units and living space are less exact than a formal count.

**Ethics:** There are no ethical issues that I’m aware of.

**Relevance:** The population and residential unit data will add another angle to my project, allowing me to determine if population, residential unit density and trends in both domains are related to certain characteristics of the rental market.

**Data Profiles**

**Immoscout24 rental data**

**Variables (red variables were removed in the cleaning process)**

* **regio1:** Region (Bundesland)
  + **Time-**invariant, structured, qualitative (nominal)
* **serviceCharge:** Additional costs such as electricity or internet
  + **Time-**invariant, structured, quantitative (continuous)
* **heatingType:** Type of heating
  + **Time-**invariant, structured, qualitative (nominal)
* **telekomTvOffer:** Is TV included? If so, which offer?
  + **Time-**invariant, structured, qualitative (nominal)
* **telekomHybridUploadSpeed:** How fast is the hybrid inter upload speed?
  + **Time-**invariant, structured, quantitative (continuous)
* **newlyConst:** Is the building newly constructed?
  + **Time-**invariant, structured, qualitative (nominal)
* **balcony:** Does the unit have a balcony?
  + **Time-**invariant, structured, qualitative (nominal)
* **picturecount:** Number of pictures attached to listing
  + **Time-**invariant, structured, quantitative (discrete)
* **pricetrend:** Price trend calculated by Immoscout
  + **Time-**variant, structured, quantitative (continuous)
* **telekomUploadSpeed:** Internet upload speed
  + **Time-**invariant, structured, quantitative (continuous)
* **totalRent:** Total rent (usually a sum of base rent, service charge and heating costs)
  + **Time-**variant, structured, quantitative (continuous)
* **yearConstructed:** Construction year
  + **Time-**invariant, structured, quantitative (continuous)
* **scoutId:** Immoscout ID
  + **Time-**invariant, structured (index variable)
* **noParkSpaces:** Number of parking spaces
  + **Time-**invariant, structured, quantitative (discrete)
* **firingTypes:** Main energy sources, separated by a colon
  + **Time-**invariant, structured, qualitative (nominal)
* **hasKitchen:** Does the unit have a kitchen?
  + **Time-**invariant, structured, qualitative (nominal)
* **geo\_bln:** Region indicator; identical to regio1
  + **Time-**invariant, structured, qualitative (nominal)
* **cellar:** Does the unit have a cellar?
  + **Time-**invariant, structured, qualitative (nominal)
* **yearConstructedRange:** Construction year range, 1 (oldest) to 9 (newest)
  + **Time-**invariant, structured, qualitative (ordinal)
* **baseRent:** Base rent without electricity and heating
  + **Time-**variant, structured, quantitative (continuous)
* **houseNumber:** House number
  + **Time-**invariant, structured, qualitative (nominal)
* **livingSpace:** Living space in square meters
  + **Time-**invariant, structured, quantitative (continuous)
* **geo\_krs:** Kreis/district (more specific than regio1); identical to regio2
  + **Time-**invariant, structured, qualitative (nominal)
* **condition:** Apartment condition
  + **Time-**variant, structured, qualitative (nominal)
* **interiorQual:** Interior quality
  + **Time-**variant, structured, qualitative (nominal)
* **petsAllowed:** Are pets allowed?
  + **Time-**invariant, structured, qualitative (nominal)
* **street:** Street name
  + **Time-**invariant, structured, qualitative (nominal)
* **streetPlain:** Street name without special characters
  + **Time-**invariant, structured, qualitative (nominal)
* **lift:** Is there an elevator?
  + **Time-**invariant, structured, qualitative (nominal)
* **baseRentRange:** Base rent range, 1 (cheapest) to 9 (most expensive)
  + **Time-**invariant, structured, qualitative (ordinal)
* **typeOfFlat:** Type of unit
  + **Time-**invariant, structured, qualitative (nominal)
* **geo\_plz:** PLZ/ZIP code
  + **Time-**invariant, structured, qualitative (nominal)
* **noRooms:** Number of rooms
  + **Time-**invariant, structured, qualitative (nominal)
* **thermalChar:** Energy needs in kWh/(m^2a); determines the energy efficiency class
  + **Time-in**variant, structured, quantitative (continuous)
* **floor:** Which floor the unit is on (note that the ground floor is 0 in Germany)
  + **Time-**invariant, structured, quantitative (discrete)
* **numberOfFloors:** Number of floors in the building
  + **Time-**invariant, structured, quantitative (discrete)
* **noRoomsRange:** Number of rooms range, 1 (fewest) to 5 (most)
  + **Time-**invariant, structured, qualitative (ordinal)
* **garden:** Does the unit have a garden?
  + **Time-**invariant, structured, qualitative (nominal)
* **livingSpaceRange:** Living space (square meters) range, 1 (smallest) to 7 (largest)
  + **Time-**invariant, structured, qualitative (ordinal)
* **regio2:** District/Kreis; identical to geo\_krs
  + **Time-**invariant, structured, qualitative (nominal)
* **regio3:** City/town
  + **Time-**invariant, structured, qualitative (nominal)
* **description:** Free text description of the unit
  + **Time-**invariant, unstructured
* **facilities:** Free text description of available facilities
  + **Time-**invariant, unstructured
* **heatingCosts:** Monthly heating costs in Euro
  + **Time-**variant, structured, quantitative (continuous)
* **energyEfficiencyClass:** Energy efficiency class
  + **Time-**invariant, structured, qualitative (ordinal)
* **lastRefurbish:** Year of last renovation
  + **Time-**invariant, structured, quantitative (continuous)
* **electricityBasePrice:** Monthly base price for electricity
  + **Time-**variant, structured, quantitative (continuous)
* **electricityKwhPrice:** Electricity price per kwh
  + **Time-**variant, structured, quantitative (continuous)
* **date:** Time of scraping
  + **Time-**invariant, structured, quantitative (continuous)

Data Cleaning

Raw data

* Rows: 268,850
* Columns: 49

Cleaned data

* Rows: 268,772
* Columns: 30

Process:

I dropped every row with more than 33% missing data (88,720 NaNs), as well as rows that contained duplicate information or information I wasn’t interested in for this analysis. I removed the following rows:

serviceCharge (6,909 NaNs, removed because of high variability with little context), telekomTvOffer (32,619 NaNs, removed due to irrelevance to my analysis), telekomHybridUploadSpeed (223,830 NaNs), telekomUploadSpeed (33,358 NaNs, irrelevance), noParkSpaces (175,798 NaNs), houseNumber (71,018 NaNs, irrelevance), geo\_bln (0 NaNs, repetitive), geo\_krs (0 NaNs, identical to regio2), interiorQual(112,665 NaNs), petsAllowed (114,573 NaNs), streetPlain (71,013 NaNs, irrelevance/repetitive), thermalChar (106,506 NaNs), numberOfFloors (97,732 NaNs), facilities (52,924 NaNs, irrelevance and difficulty involved in analysis), heatingCosts (183,332 NaNs), energyEfficiencyClass (191,063 NaNs), lastRefurbish (188,139 NaNs), electricityBasePrice (222,004 NaNs), electricityKwhPrice (222,004 NaNs)

The removals left me with 30 rows, not including the index.

Here’s a summary of other changes I made during the cleaning process:

* **totalRent:** Replaced 14 entries where the rent was inordinately high (over 10,000 Euro) and the description didn’t indicate the rental is a luxury unit. – imputed median rent in the region. Replaced 236 zeroes with the value from the row’s baseRent column. Dropped 78 rows with totalRent values of less than 100.
* **baseRent:** Replaced 89 zeroes with the value from the row’s totalRent column. Replaced six entries where base rent was inordinately high (over 10,000 Euro) and the description didn’t indicate the rental is a luxury unit – imputed median rent in the region.
* **yearConstructed:** Corrected five incorrect dates (too far in the future) by inferring construction date from description and scraping date. Replaced 87 incorrect construction years (1111, looks to be a placeholder) with column mean.
* **livingSpace:** Replaced eight inordinately high livingSpace values with mean for the given noRoomsRange. Corrected one value that included the correct value in the description. Corrected 9 values where livingSpace was swapped with noRooms. Replaced 71 zeroes with NaN.
* **noRooms:** Corrected/replaced 30 inordinately high noRooms values with the value stated in the description or the median value for the given livingSpaceRange.
* **floor:** Corrected/replaced 31 suspiciously high floor values with the value stated in the description or the median value in the given region.
* **scoutId:** Changed data type to string.
* **date:** Changed to datetime data type.

I also addressed mixed-type data in five columns (firingTypes, heatingType, condition, typeOfFlat, description) by changing the data type to string.

Descriptive Statistics (post-cleaning)

(tables are on next page)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | baseRent | totalRent | pricetrend | livingSpace | picturecount | yearConstructed |
| count | 268772 | 228223 | 266940 | 268701 | 268772 | 211765 |
| mean | 647.66 | 814.82 | 3.39 | 73.46 | 9.79 | 1966.75 |
| std | 505.61 | 562.47 | 1.96 | 32.16 | 6.41 | 43.68 |
| min | 1 | 100 | -12.33 | 5 | 0 | 1000 |
| 25% | 338 | 470 | 2 | 54 | 6 | 1950 |
| 50% | 490 | 650 | 3.39 | 67.28 | 9 | 1973 |
| 75% | 799 | 985 | 4.57 | 86.91 | 13 | 1996 |
| max | 20000 | 26500 | 14.92 | 1717.74 | 121 | 2022 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | yearConstructed | scoutId | noRooms | floor | baseRentRange | noRoomsRange | livingSpaceRange | yearConstructedRange |
| count | 211765 | 268772 | 268772 | 217493 | 268772 | 268772 | 268772 | 211765 |
| mean | 1966.75 | 106970106.29 | 2.63 | 2.10 | 3.77 | 2.57 | 3.07 | 3.71 |
| std | 43.68 | 12500132.87 | 0.98 | 1.66 | 2.21 | 0.94 | 1.41 | 2.74 |
| min | 1000 | 28871743 | 1 | -1 | 1 | 1 | 1 | 1 |
| 25% | 1950 | 106691240.5 | 2 | 1 | 2 | 2 | 2 | 1 |
| 50% | 1973 | 111158381.5 | 3 | 2 | 3 | 3 | 3 | 3 |
| 75% | 1996 | 113768747.5 | 3 | 3 | 5 | 3 | 4 | 5 |
| max | 2022 | 115711743 | 18 | 41 | 9 | 5 | 7 | 9 |

**Statistisches Bundesamt population and apartment data**

**Variables**

* **Year: Year of data collection**
  + **Time-variant, structured, quantitative (continuous)**
* **State: Federal state**
  + **Time-**invariant, structured, qualitative (nominal)
* **Apartments: Number of apartments (rental units) in the region**
  + **Time-variant, structured, quantitative (discrete)**
* **Living space: Living space of all apartments combined in units of 1,000 square meters**
  + **Time-variant, structured, quantitative (continuous)**
* **Population: Total population of the region**
  + **Time-variant, structured, quantitative (discrete)**
* **People/apt: Number of residents divided by number of apartments**
  + **Time-variant, structured, quantitative (continuous)**
* **Space/person: Living space in square meters divided by number of residents**
  + **Time-variant, structured, quantitative (continuous)**

Data Cleaning

Raw data

* Rows: 374
* Columns: 7

Cleaned data (unchanged)

* Rows: 374
* Columns: 7

Process:

The dataset is small and needed minimal cleaning. The only changes I made were in the state column, where I changed dashes (-) to underscores (\_) to match the formatting of the Immoscout24 dataset.

Descriptive Statistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | year | apartments | living space | population | people/apt | space/person |
| count | 374 | 374 | 374 | 374 | 374 | 374 |
| mean | 2010.5 | 4648866 | 413646.9 | 9666740 | 2.023561 | 42.46926 |
| std | 6.352787 | 8982525 | 802638.2 | 18716557 | 0.133175 | 4.139222 |
| min | 2000 | 337002 | 25934 | 652182 | 1.718245 | 33.5974 |
| 25% | 2005 | 1132627 | 87430 | 2157101 | 1.922368 | 39.14743 |
| 50% | 2010.5 | 1856260 | 135362 | 3426752 | 2.032761 | 42.34583 |
| 75% | 2016 | 3798064 | 388859.5 | 7990733 | 2.12598 | 44.97164 |
| max | 2021 | 41660773 | 3841439 | 83237124 | 2.291101 | 53.6765 |

**Questions to Explore**

How do total and base rent vary based on geography?

Has rent increased over time?

Is rent higher in areas with faster-growing population, higher population density and less living space relative to population?

How does the number of pictures attached to an advertisement relate to the rental price?

Are units with kitchens more expensive than units without kitchens?

Are kitchens more common in some areas than others?

Do electricity and heating type have a correlation with rental prices?

Does the unit floor have a correlation with price?

Does the presence of a balcony and/or garden have a correlation with price?

Are older units more or less affordable than newer units?

How do refurbishment and apartment type relate to rental price?

Does the completeness of the listing correlate with rental price?