How large is the current rent price bubble really?

A descriptive study of the German housing market

Computational Social Science: Seminar Paper

Marc Lüttecke*

Department of Economics

marc.luettecke@uni-konstanz.de

Student ID: 910426

Advisor: Dr. Karsten Donnay

Abstract

The question of drivers for human behavior is complex and finds its epitome in financial- and, consequently, housing-markets through the simultaneous interaction of often millions of agents. This paper allows a first glimpse into exciting interactions between commonly available economic indicators, such as unemployment-rates, crime-rates, purchasing power indexes and the rent prices of the most influential German metropolitan areas. A data validity section establishes legitimacy of the scraped data-set by investigating if subjective price overestimation persists, while an interactive visualization encourages the reader to take an active role throughout the paper to observe surprising effects: often, only the absolute values of rent prices appear astonishingly high in *pricey* cities, but a higher local purchasing power index, allows consumers to own more money in the first place. Low crime- and unemployment-rates demand a price premium for rent, which establishes an unexpected subset of cities to be overpriced in relative terms. Some city-officials seem aware of the excess demand/supply discrepancy and act proactively, which might offer predictive power for prospective rent bubbles.

1 Introduction

1.1 Motivation and research gap

Economic prosperity and ever-increasing outreaches of globalization have led to an amount of wealth and scientific research unknown to humanity at any other time. The expanding wealth of countries has also given rise to an unequal distribution of this wealth among nations and individuals[1].

*Lüttecke is a Masters student in *Social and Economic Data Science* at the University of Konstanz. This research project represents the seminar paper to Karsten Donnay's class *Computational Social Science*. The author would like to explicitly express his gratitude for the forum to conduct this research endeavor, as well as the constructive feedback from the cohort of seminar members.

1 See also: https://www.theguardian.com/news/2019/sep/09/inequality-is-it-rising-and-can-we-reverse-it

One of the main re-occurring expenses in most people's life is the monthly rent payments[2]. While rent prices generally align with general indicators of purchasing power or wealth, often, certain regions experience a kind of bubble[3], which the meltdown of the financial system in 2007[4] most prominently depicted. The economic outreach and consequences caused century-long established financial service providers to declare bankruptcy and shook entire nations to its core with a consequent need to governmental bail-outs.

The explanations for rent price bubbles are diverse and range from intuitive (f.ex. attractive regions lead to inflated prices due to high-demand, or at least speculative interests) to more exotic (as in Joebges et al. [17], which claims that financial innovation can disrupt housing markets and let prices deviate from their long-term equilibrium[5]). While the real reasons still seem opaque, the economic and social research community was able to pinpoint certain instigators, which lead to an increase in attractiveness in regions and, in response, in market prices. Common themes are crime rates (Schwartz et al. [29], Bishop and Murphy [3], Harcourt [12]), unemployment rates (Branch et al. [5], Chao and Eden [7], Oswald [24]), purchasing power[6] (Taylor and Taylor [31], Inklaar and Timmer [15]), and inflation[7] (Rubens et al. [28], Hoesli [14], Anari and Kolari [1]). Since these papers often deal with the interactions in isolation and define a very narrow market, such as the New York (Schwartz et al. [29]), or the Los Angeles (Harcourt [12]) housing market, wider generalizability is mostly lacking from current literature. Social scientists describe these systems as *complex systems* since many moving parts are interacting in ways that, especially given the irrational and dynamic human component, are hard to replicate, isolate, and therefore investigate (see Helbing [13]).

1.2 Main contributions

Given that the effect is highly interdependent, this paper strives for a more modest and intuitive research goal: a descriptive study of effects on the German rental price market that hints at possible interactions. While rigorous statistical analysis is vital for the identification of correlation and potentially even causal effects, it often lacks the layman interpretability that descriptive studies provide. The explanatory bar is intentional set low: to include all the endogeneous effects that would disqualify any causal interpretation. Therefore, the main research contributions are:

1. Provision of a large dataset of the 14 largest metropolitan areas in Germany for rental prices of one- and three-room apartments through the largest online rental market website[8].

2. Comparison of these city prices (and for verification purposes about 80 more) with manually listed cost-of-living estimates on the largest site that I am aware of to publicly compare the cost of common consumer goods

2 See also: https://www.study-in-germany.de/en/plan-your-stay/money-and-costs/cost-of-living_28220.php.
3 A *financial bubble* represents a sharp increase in the price of an asset, which is not reflected by its actual value anymore.
4 https://www.investopedia.com/articles/economics/09/financial-crisis-review.asp.
5 Also referred to as *Market Equilibrium*, a state, in which economic forces are balanced and, by themselves, do not deviate further.
6 Financial ability to purchase goods.
7 A measure of price increases of a predefined basket of goods, often consumer goods, to measure how much money still really buys.
8 Source: https://www.immobilienscout24.de/

for cities worldwide[9] [10].

3. Matching of the local city prices with often-researched economic indicators for rental prices, such as local purchasing power, crime rates, and unemployment rate.

4. Evaluation of the respective response of cities to these insights, i.e., are urban areas building more apartments that suffer from high rental prices and disproportional economic indicators.

The structure of the remainder of the paper is as follows: section 2 will provide an overview over the literature that addresses rental price bubbles, section 3 explains in detail the research design of this study including data retrieval and research hypotheses, section 4 illustrates the results, section 5 enlists possible model extensions and section 6 concludes.

2 Related work

The body of literature for rent prices is rich and exceeds the extend of this research endeavor. This section outlines the most prominent papers on the general groundwork for the approach chosen, namely *complex systems in social sciences* before breaking up into the individual components that could identify rent prices, namely: unemployment rate, crime rate and purchasing power in interaction with rent prices, as well as education as an honorable mention.

2.1 Rent prices as complex social systems

As the authors discuss in Helbing [13], the problems of *"modeling socio-economic systems"* arises from many sources. The one most particular to the rent price modelings, are (see Helbing [13] for a full list of potential complications):

- *Frequently there is no ensemble of equivalent systems, but just one realization (one human history).*
- *The non-linear and/or network dependence of many variables lead to complex dynamics and structures, and sometimes paradoxical effects.*
- *Factors such as a significant degree of randomness and heterogeneity, memory, anticipation, decision-making, communication, consciousness, and the relevance of intentions and individual interpretations complicate the analysis and modeling a lot.*
- *In particular, social systems are influenced by normative and moral issues, which are variable.*

Jenkins [16] provides a vivid example, which relates rent-control proposals to complex system theory. It offers a systematic review of economists' opinions on rent control, which for the longest time, stood as a first exposure to an example of price-ceiling-theory and inefficiency of markets for intro to economics students.

9 Source: https://www.numbeo.com/cost-of-living/

10 Another alternative represents the widely available *Mietspiegel* metric in Germany, which offers average rent prices per m^2. The decision against it stems from two reasons: (1) most Mietspiegel figures are also locally scrapped rent prices, which does not differ to the approach of scraping ImmoScout24 and (2) the exact method of deriving the rent averages, including city-center-distance, time period, frequency, selection bias, etc., are not displayed. See for example: https://www.wohnungsboerse.net/mietspiegel-mietpreise

Łaszek [21] details an alternative body of work, which builds on consumer theory [11] to mimic decision-making in the housing market, the derivation of a complex demand function, and the subsequent impact on the overall housing market.

Marsh and Gibb [23] grants a third example of economic theory to include housing market decisions. The authors explain that the classical models, namely expected utility theory[12], for expectation formation is ill-suited for the task of predicting house and rent expenses. An alternative route using the tool-set of behavioral economists [13] is stated as more appropriate for the task.

This brief extract aims to clarify one central truth. Rent prices are as difficult to predict, as numerous other problems arising in many social sciences in recent decades. Easily quantifiable statistics and mathematical models even fail to predict individual human behavior accurately and consequently often also fail for the behavior of masses in financial or housing markets. The rise of behavioral economics (with its epitome of the Nobel price acknowledgment to Richard Thaler in 2017, Appelbaum [2]) is a prominent symptom of this fact. To limit the confusion and wrong assumptions made by over-extending their research findings, many authors limit their question to one specific factor to change and a *ceteris paribus*[14] assumption for the rest.

2.2 Unemployment rate and rent

Rahman and Mustafa [27] details a study investigating the effect of US unemployment rates on the US housing market index. Advanced statistical time-series techniques, such as vector error-correction model (VECM)[15], KPSS[16] tests to investigate stationarity in the time series, and Johansen-Juselius co-integration procedures[17]. They find evidence of long-run convergences of the two time-series, meaning that the two-time series impact one another. However, the effect seems to be reaching both ways: a clear one-way causal relationship is difficult to establish, which hints at the complex nature of the rent-price behavior that this research endeavor tries to address. While Rahman and Mustafa

11 *Consumer theory* refers to a branch of Microeconomics, which uses money as the primary vehicle of quantitative decision making. It researches individuals' decision-making, given certain budget constraints to derive their utility function.

12 Under *Expected Utility Theory*, an economic agent quantifies its utility function, a formulated model to describe her happiness, including all factors that impact it, under different expectations for future values of the parameter that formulate her utility.

13 *Behavioral Economics* is a branch of modern economics that tries to close some of the cognitive gaps in classical utility-based, rational, full-information economics, by incorporating agents' irrationality as a factor into quantifiable models. It, therefore, strongly overlaps with branches and theories of psychology.

14 Latin for *other things equal*, meaning for mostly research or thought experiments, one factor is changed to isolate its effect, while all others remain the same, which allows for a more causal interpretation of effects.

15 An error correction model (ECM), belongs to a category of time-series models most commonly used for data, for which the underlying variables have a long-run stochastic trend, also known as cointegration. ECMs are a theoretically-driven approach useful for estimating both short-term and long-term effects of one-time series on another. The term error-correction relates to the fact that the last period's deviation from a long-run equilibrium, the error, influences its short-run dynamics - https://en.wikipedia.org/wiki/Error_correction_model.

16 In econometrics, *Kwiatkowski–Phillips–Schmidt–Shin (KPSS)* tests are used for testing a null hypothesis that an observable time series is stationary around a deterministic trend (i.e. trend-stationary) against the alternative of a unit root - https://en.wikipedia.org/wiki/KPSS_test.

17 In statistics, the *Johansen test*, is a procedure for testing cointegration of several, say k, I(1) time series - https://en.wikipedia.org/wiki/Johansen_test.

[27] establishes statistical significance, the authors agree that economic effects remain minuscule.

A more classical paper of Oswald [24] prepares a question that aims in the opposite direction: how does home-ownership impact unemployment rates? Anecdotal evidence of Spain's high and Switzerland's low home-ownership hints, in combination with their high and low unemployment rate, respectively, at an interaction of low worker-mobility and consequently higher unemployment rates. While the paper intentionally omits excessive statistical rigor, the interpretability of the authors intepret correlation as causation to a questionable degree. This research project, with potentially overstated results, indicates once again the complex interdependence of the individual factors with the housing market.

Referencing the last housing market bubble and the relationship of housing prices with equity prices, after controlling for unemployment, Borio and McGuire [4] reveals a large international study of major industrialized countries and a long time-series over three decades. The authors find a cushioning relationship of the interest rate market through an inelastic real-estate market, which highlights the real-estate market as an investment alternative which is less volatile than comparable financial markets.

2.3 Crime rate and rent

Crime rates are often listed as the main indicator of well-being in a city and might indicate structural problems if unproportionally high. A vast literature from criminology to economics deals with the impact of house prices on crime, or for this paper more interestingly, the reverse effects. Tita et al. [32] explains how crime rates, differentiated between violent crime and property crime cases, affect housing prices. The new granularity of the data and a model design, which accounts for the inaccuracy of reported crime data, extends the literature and establishes crime rates as an early prediction mechanism for community transitions. Crime is capitalized differently for the distinct social classes of poor, middle and upper-class citizens, while violent crime imparts the highest cost.

Bishop and Murphy [3] addresses the existing literature of hedonic models[18], which generally include a myopic[19] decision-maker to allow for more rational models for housing decisions, which most consumers have to endure for a long time. They find that agents are willing to pay a premium of $ 472 annually for a 10% reduction in violent crime.

In a more dated and targeted study, Schwartz et al. [29] asks whether decreasing crime-rates generally drive New York city's post-1994 housing boom, which expresses the popular opinion of the time, or if other factors, such as low-income housing subsidies, or the quality of public schooling drove the increase in property value. They close with the consensus that lower crime rates account for about a third of the price boom while failing to include a more specific figure to generalize to other cities, for example, through a cross-sectional study with interpretable coefficients. The approach implies a more exemplary study, but provides a generally comparable endeavor to this paper, asking: how come rent prices have changed so much. Once again, the complexity of asking if either of four

18 *Hedonic pricing* refers to models that identify price factors according to the premise that price is determined both by internal characteristics of the good sold and external factors affecting it - https://www.investopedia.com/terms/h/hedonicpricing.asp.

19 Agents in economics are often perfectly rational. *Myopic* decision-makers loosen this definition by over-weighting the near future and under-weighting long-term effects.

economic factors or any combination of them drive rent prices in a cross-section of 15 cities becomes clear.

Buonanno et al. [6] illustrates an exciting approach to account for the perception of crime rates, while omitting statistical rigor of valid crime rates. Their approach includes a hedonic pricing model, incorporating an OLS[20] and a quantile regression[21] model to estimate house price developments in Barcelona according to perceived security (data stems from a victimization survey). After controlling for district characteristics, such as local public spending and immigration and accounting for potential endogeneity concerns, they perceive a price premium of 0.57 % for a one-standard-deviation increase in subjective security, while unsafer districts lie on average 1.27% lower. This paper explains graphically how pure statistical measurements might not suffice to explain rent price behavior, but indicate how more abundant data of actual psychological perception can enrich the analysis.

2.4 Purchasing power and rent

The most obvious factor construing rent price models is to align housing prices with purchasing power information. If every single dollar buys additional goods, the scale of rent price comparisons is skewed and not universally comparable. Vastmans and Buyst [33] demonstrates this concept by researching Belgian housing prices and their determinant by either relative income or purchasing power for housing. Their findings suggest a more stable price-to-purchasing power ratio than the respective price-to-income counterpart. The interest rate elasticity[22] of the enhanced model is higher than in previous work.

Since purchasing power and inflation closely relate, more classical studies of the inflation-protection literature exist as in the form of Rubens et al. [28]. This paper of 1989 investigates if real estate properties define a valid hedging[23] vehicle for investment purposes and, therefore, a valuable addition to an investment portfolio. In a follow-up study Hoesli [14] extents the idea to the Swiss case. Both studies validate real estate as a promising hedge against inflation, while Hoesli [14] even expresses that housing prices might be a valuable predictor of inflation (especially the unexpected portion of it). Anari and Kolari [1] exchanges house price returns with actual house prices and inflation index information with cost of goods excluding rental costs and confirm the hedging ability of house prices through an ARDL model[24] and a consistent Fisher coefficient larger than 1.

Housing as a major descriptor exerts the most prominent position within a debate of housing prices, and purchasing power parity (PPP)[25], which has been an extensive research focus especially since the breakdown of Bretton

20 *Ordinary Least Squares* Regression provides the standard linear regression model, under which the sum of least squares of the difference between the estimated line to the actual data points serves as the objective function to be minimized.

21 In comparison to *OlS*, which estimates the mean impact of the independent on the dependent variable, a *quantile regression model* estimates the conditional median, or other quantiles of the response variable.

22 The concept of *elasticity* in economics describes how demand for specific good changes for a 1-unit change in price, therefore, a determinant of how easily substituted a specific good is.

23 *Heding* refers to taking an offsetting financial investment position to alleviate a given risk, ideally because the two assets are negatively correlated and, therefore, counteract each other.

24 *Autoregressive distributed lag* models refer to a common technique in time-series modeling, in which a past value of itself partly determines the current value of a variable.

25 *Purchasing power parity* is a theory that compares prices of different areas using a common good or goods to contrast the real purchasing power between different currencies - https://en.wikipedia.org/wiki/Purchasing_power_parity.

Woods[26] in the early 1970's. This development also highlights the interdependence of the housing prices with the purchasing power index itself, which might lead to complicated reverse-causality problems. Section 5 provides a more detailed discussion and potential alleviations.

2.5 Education and rent

Other indicators explore the relationship and intrinsic nature of rent prices as well. An honorable mention includes *education* (Dubin and Goodman [10], Fack and Grenet [11], Wen et al. [35]) as a driving factor. This research project intentionally excludes school performance as a predictor for rental prices, due to multiple reasons:

- School performance data offers high heterogeneity on district and even individual school-level. Different types of schools might reflect different dynamics, such as a *Hauptschule* right next to a *Gymnasium*. Which of the two would be the driving factor of rent prices close-by? More granular models, potentially limited to a single city, might be a better fit for researching this interaction.
- The different states incorporate distinct school systems as unique as varying lengths of schooling, mandatory exit exams after 10^{th} grade, or idiosyncratic number of *Leistungskurse*. All this factors would convolute the results and weaken interpretability (which is one of the main contributions of the country-wide visualization).
- The scraped data itself does not include intercity information. The mapping to exact addresses and to, therefore, allow a mapping of within-neighborhood dynamics, provides an interesting extension.

3 Research design

This section details the explicit research design of this study, explaining the process of data retrieval, as well as the entire data pipeline to allow for the data verification and the data visualization results.

The specific research questions to be asked, as well as the stated hypotheses, are:

- **RQ1 :** Are rent prices subjectively perceived differently than their actual development suggests?

  Hypothesis: Rental prices are perceived higher than they are, given the widely-shared media exposure to local rental price spikes.

- **RQ2 :** How do economic indicators relate to city-based rental price developments? Is this impacted by apartment size, or proximity to the city center?

  Hypothesis: Economic indicators, such as purchasing power or crime rates, offer a reliable indicator of the quality of living and, consequently, the prices charged in cities.

26 The *Bretton Woods* system of monetary management established the rules for commercial and financial relations among the United States, Canada, Western European countries and was the first example of a fully negotiated monetary order intended to govern monetary relations among independent states. On 15 August 1971, the United States unilaterally terminated convertibility of the US dollar to gold, effectively bringing the Bretton Woods system to an end - https://en.wikipedia.org/wiki/Bretton_Woods_system.

- **RQ3 :** Are city-officials aware and try to alleviate the spike in rental prices?

  Hypothesis: City administrations are addressing the issue, but the effect of high prices is not sufficiently weakened.

The core question of this paper lies with RQ2. This identification is in line with the research gap identified in subsection 1.1, which leads to many practical applications, such as the recognition of relevant factors that determine high rent prices, possible alleviation through also increasing purchasing power or the simple explanation that renters pay a premium for increased quality of living in a particular region (such as low crime rates, or low unemployment rates). However, to answer this enigma, which in the past has only been addressed superficially (see section 2), a thorough verification of data quality is vital.

RQ1 addresses this endeavor to, namely, compare the scraped data with a widely available online comparison website for price experiences. *Numbeo* serves as a comparative-tool to list relative purchasing power in certain regions, for consumers to check if, for example, a move to a new city is lucrative, given prospective income expectations of a new job. The website states that "*Numbeo is the world's largest database of user-contributed data about cities and countries worldwide. Numbeo provides current and timely information on world living conditions, including the cost of living, housing indicators, health care, traffic, crime and pollution. - 5,783,185 prices in 9,246 cities entered by 493,811 contributors*"[27], which classifies it as a valid benchmark for the scraped dataset.

RQ2 lies at the heart of the research project to interactively allow the user to experience potential discrepancies in rent prices with widely available economic indicators. What determines rent prices and how do even multiple factors interact with local rent prices, is a complex questions that discerns precludes many statistical methods, which require large data-sets to establish significance.

RQ3 concludes the descriptive nature of the research question to ask, more specifically, if cities are aware of potential problems and if they react in a manner that is in line with the extent of the discrepancy that this paper identifies. So, if a city is overpriced, relative to economic indicators, are cities building new apartments at a higher rate than in other cities? Is the new building rate of apartments unrelated to these measures and entirely exogenous to these criteria?

The data pipeline, illustrated in Figure 1, depicts the software usage of this project. I scraped publicly available data from ImmoScout24 via a Python script, processed it into an appropriate format, and uploaded it to a cloud version of a MongoDB[28]. From there, all future analyses directly download the data, which allows for a continuous expansion of the database for analysis. RQ1 is then addressed in Python, utilizing standard statistics packages[29], which allows for the comparison of distributions of variables (between the scraped data and the information entered from Numbeo). RQ2 builds on an interactive visualization, built with the leaflet package[30] package in *R*. Additional

27 For all major metropolitan areas, individual contributor and entry numbers are listed online (but are generally in the hundreds or thousands). The data retrieval horizon ranges over the last 18 months. Exemplary source bottom of: https://www.numbeo.com/cost-of-living/in/Berlin

28 *Atlas* is the official online solution offered by MongoDB and can be accessed via a free account under https://cloud.mongodb.com/

29 https://docs.scipy.org/doc/scipy/reference/stats.html

30 *Leaflet* is a Javascript-based package that allows different software languages to build interactive maps. They contain a large database of country and city maps as blueprints. Source: https://leafletjs.com/

information for economic indicators are partly entered automatically, but due to its small nature (sometimes we have only unemployment rates on state (*Bundesland*) level), partly entered manually from different sources (see subsection 3.1 for details).

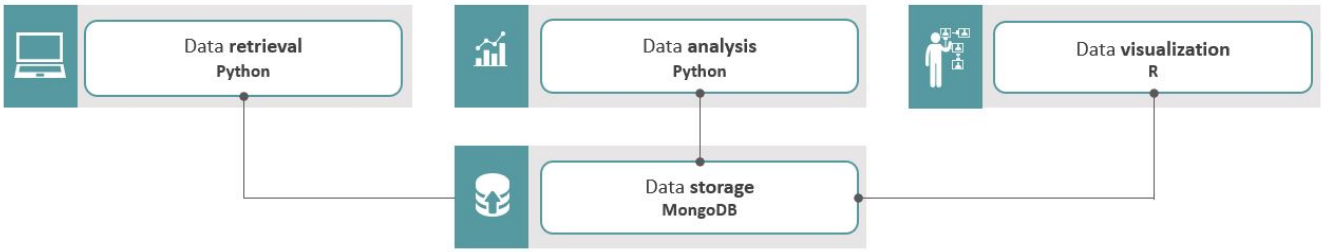


Figure 1: Illustration of the data pipeline process, highlighting the multi-software usage for specific project stages. Here, every step interacts with a cloud version of MongoDB to receive, analyze, or visualize the most recently updated dataset to allow for a dynamic interaction of the various analysis components. The process step can be found at the top, while the respective software is listed below, both printed in ***bold****.*

3.1 Data retrieval

One of the main contributions of this study is the extensiveness of the data and its dynamic growth that is continuously enhanced past the final write-up of this paper. Since Numbeo lists the cost of living information for 67 cities in Germany, the scraper for the data to answer RQ1 (the data validity task) searched for all information of rent prices off ImmoScout24 over a range of dates for 1- and 3-bedroom apartments. This results in an overall sample size (as of the time of writing) of 44,422 documents in the MongoDB, which is averaged for the specific categories listed on Numbeo: 1-bedroom city-center, 3-bedrooms city-center, 1-bedroom outside of city-center, 3-bedrooms outside of city-center. Since Numbeo does not specify, what exactly qualifies as *city-center*, the scraped documents include all offers within 15 kilometers from the actual city center, while a cutoff of five kilometers differentiates between *city-center* or *outside of city-center*.

Due to practical reasons (scraping all offers of 67 German cities takes about two nights) and the lack of city-level data on some of the economic indicators, the research endeavor of RQ2 (the comparison of rent prices to common economic indicators), limits the approach to the 14 largest metropolitan areas in Germany. The entire dataset for this subset of cities contains (as of writing) 24,594 entries, for which the rent prices individually build the averages congruently to the method for the list of 67 cities depicted above.

The economic data is freely available online and are available under the following sources:

- Population: Wikipedia[31]
- Crime rates: Bundeskriminalamt Deutschland[32]
- Purchasing power: Study by data service provider - *Growth From Knowledge*[33]

31 Source: https://en.wikipedia.org/wiki/List_of_cities_in_Germany_by_population
32 Source: https://www.bka.de/DE/AktuelleInformationen/StatistikenLagebilder/PolizeilicheKriminalstatistik/PKS2018/pks2018_node.html
33 Source: https://www.gfk.com/insights/press-release/purchasing-power-germany-2018/

- Unemployment rate: Statista - Federal German Statistics Bureau[34]
- New apartment rate: Statista - Federal German Statistics Bureau[35]

4 and 5 summarizes the resulting data from scraping in section 4. The collection time frame runs from November to December 2019. Exact frequencies of scraping efforts remain secondary here since MongoDB allows for the unique filtering of duplicates. Meaning, a total of 100 records does not naturally account for ten scraping sessions of ten records each but depend strongly on the new listings posted on ImmoScout24 for a given day. Generally, every day, all new postings expand the analysis to account for a total of 24,594 individual listings for the 14 cities. The extended data set was scraped fewer times since every run through the 67 cities takes about two-three days. The time frame lies within November and December, 2019.

A user-defined API to scrape Numbeo's data allows to address RQ1 and compare the rent prices off ImmoScout24 with the proclaimed information entered on Numbeo[36]. This Python script enters the information of all German cities into a reader-friendly JSON [37] format, which further incorporates the data comparison, which section 3.2 explains.

3.2 Methodology

The methodology to analyze the data includes two main parts, namely the section to address RQ1 and RQ2/RQ3. To answer RQ1, whether the scraped data is valid and if people generally overestimate their cost of living on Numbeo, the paper compares the general distributions of scraped data off Numbeo with the scraped data off ImmoScout24.

The individually scrapped records of ImmoScout24 data is first classified into 1- and 3-bedroom apartments and filtered through the metric "distance to city center" that was freely available for every posting on ImmoScout24 into (1) less than 5 km (city-center) and (2) more than 5km and less than 15km (outskirts). This classification stems from the categorization inherent on Numbeo itself, which follows the same pattern (without specifying the exact definition of 'city-center' and 'away from city-center'). Therefore the entire data set combines entries for 67 cities with four ImmoScout24 and four Numbeo columns (city-center/outskirts; 1-bedroom/3-bedrooms).

The test statistic to compare these four distributions is the *Kolmogorov-Smirnov* test (from now on KS test), and section 4.2 reports the relevant F-statistics. The statistical properties of the KS test are detailed in a quick aside in section 3.2.1.

While the statistical rigor of a test-statistic, to answer RQ1 drives very quantifiable results, the aspired intuition of RQ2's and RQ3's methods are less statistically-verifiable. As detailed before, since the problem of building a descriptive study for the housing market and its main driving factors is a socially complex problem, this paper is not aiming for definite answers. However, the insights that can be build from allowing interacting effects, without

34 Source: https://de.statista.com/statistik/daten/studie/36651/umfrage/arbeitslosenquote-in-deutschland-nach-bundeslaendern/.
35 Source: https://de.statista.com/statistik/daten/studie/1298/umfrage/wohnungsbestand-nach-bundeslaendern/
36 Source: https://github.com/mounicmadiraju/scrape.
37 *JavaScript Object Notation* is an open-standard file format or data interchange format that uses human-readable text to transmit data objects consisting of attribute–value pairs and array data types (or any other serializable value). Source: https://en.wikipedia.org/wiki/JSON

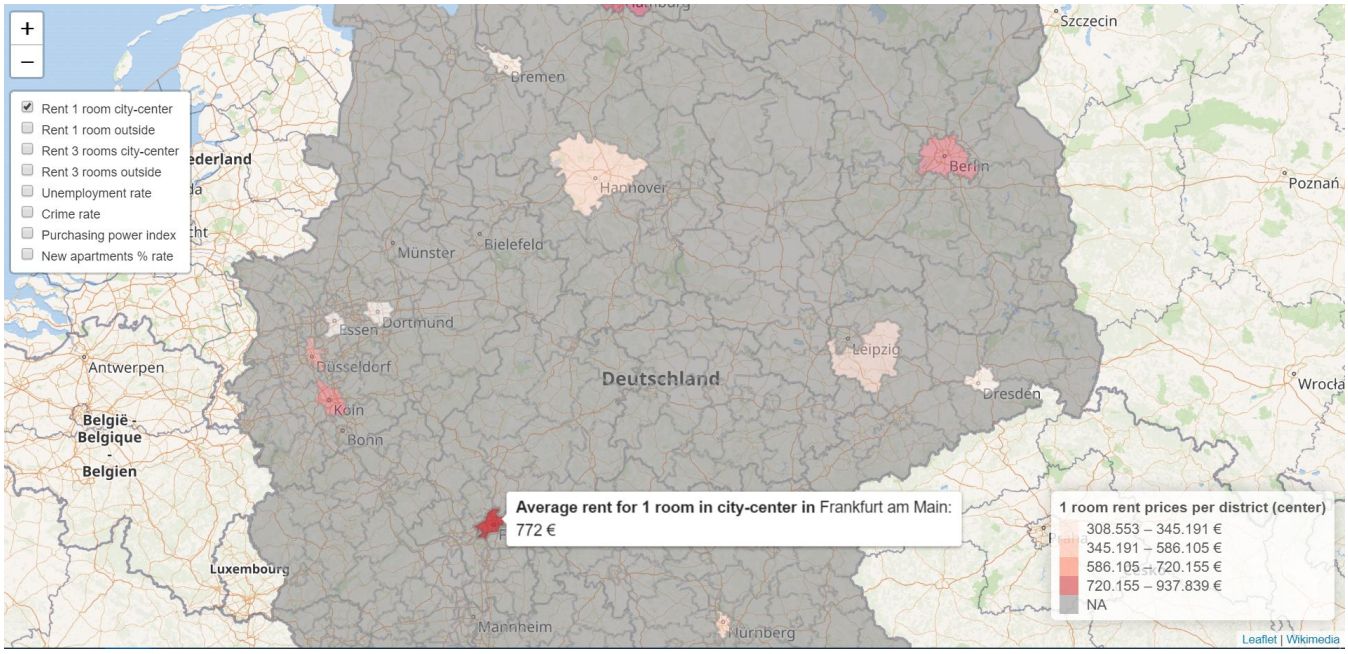


Figure 2: Illustration of interactive map to visualize the interaction of economic indicators with the rent prices to find main drivers of the German housing market. The .html file allows for hovering, zooming, multi-layer coloring and dense information provision.

a special focus on mathematical rigor, usually *limiting* descriptive studies, the paper's main contribution lies with assemble a starting block for future work.

Therefore, rent prices are mapped to an interactive map of Germany with its metropolitan regions and hovering over the individual cities, gives an overview of primary statistics, such as population. By appending layers to the map, a user can highlight rent prices for the four categories explained above and boost intuition for the renting landscape of 14 major cities in Germany. Additionally, filters of economic indicators for various color-schemes to overlay the existing rent filters, grant an effect of visualizing counter-acting effects, if for example, a high rent price (very red) is also experiencing high purchasing power (very blue) to almost lead to a neutral beige coloring of the city. This more descriptive than mathematical approach affords further combinations than usually possible in a study of 14 observations. Also, the cumbersome table of information (see a list of rent prices for the broad set of 67 cities and try to find the 14 largest cities in it in 5) effectively narrows down to main effects to a visualization allowing hovering, interaction, zooming and informative coloring.

Figure 2 provides a first example of the resulting map without any further result discussion.

This illustration provides insights into the economic indicators that could interact with rent prices of different categories (RQ2), but also into an answer to the research question if city-officials are actively addressing rent bubbles by building new apartments (RQ3). The final layer, depicted in Figure 2 on the left, offers insight into this question. Results are further addressed in section 4.

3.2.1 Aside: Kolmogorov-Smirnov Test

The used two-sample Kolmogorov-Smirnov test, serves as a test-statistic to identify if two samples derive from the same distribution. A main premise is that the distributions in question are continuous[38]. The KS test establishes a distance metric between the two samples with a null hypothesis that both samples are drawn from the same distribution. This non-parametric[39] method is both sensitive to differences in the location and shape of the empirical cumulative distribution functions of the two samples.

The KS statistic is defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}\left(X_i\right)$$

where $I_{[-\infty,x]}(X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ x and equal to 0 otherwise.

The actual test-statistic is then derived as:

$$D_n = \sup_x |F_n(x) - F(x)|$$

where the supremum[40] (here sup_x) converges to 0 *almost surely*[41] in the limit for a large enough sample size n, if the samples originate from the same distribution.

4 Results

4.1 Descriptive data exploration

Before the more advanced methods of a quantitative comparison of the underlying distribution will offer insights of the statistical relationship between the scraped data sets, less rigorous and more descriptive statistics will allow a better intuition of the underlying mechanics and expected interaction of the variables. It is important to notice that the information in Table 1[42] only builds intuition, since no confidence intervals, or R^2 (or other quantifiers of statistical reliability) are listed.

Figure 3 allows a first hint at the distribution comparison between the ImmoScout24 and Numbeo data sets. With the help of a kernel density smoothing, inherent differences are captured and show that on average Numbeo prices are higher, and slightly more dispersed. This tendency might hint at less congruence of the information entered, and potentially less of a selection bias (people do not as closely agree on the prices entered as for the ImmoScout24 data

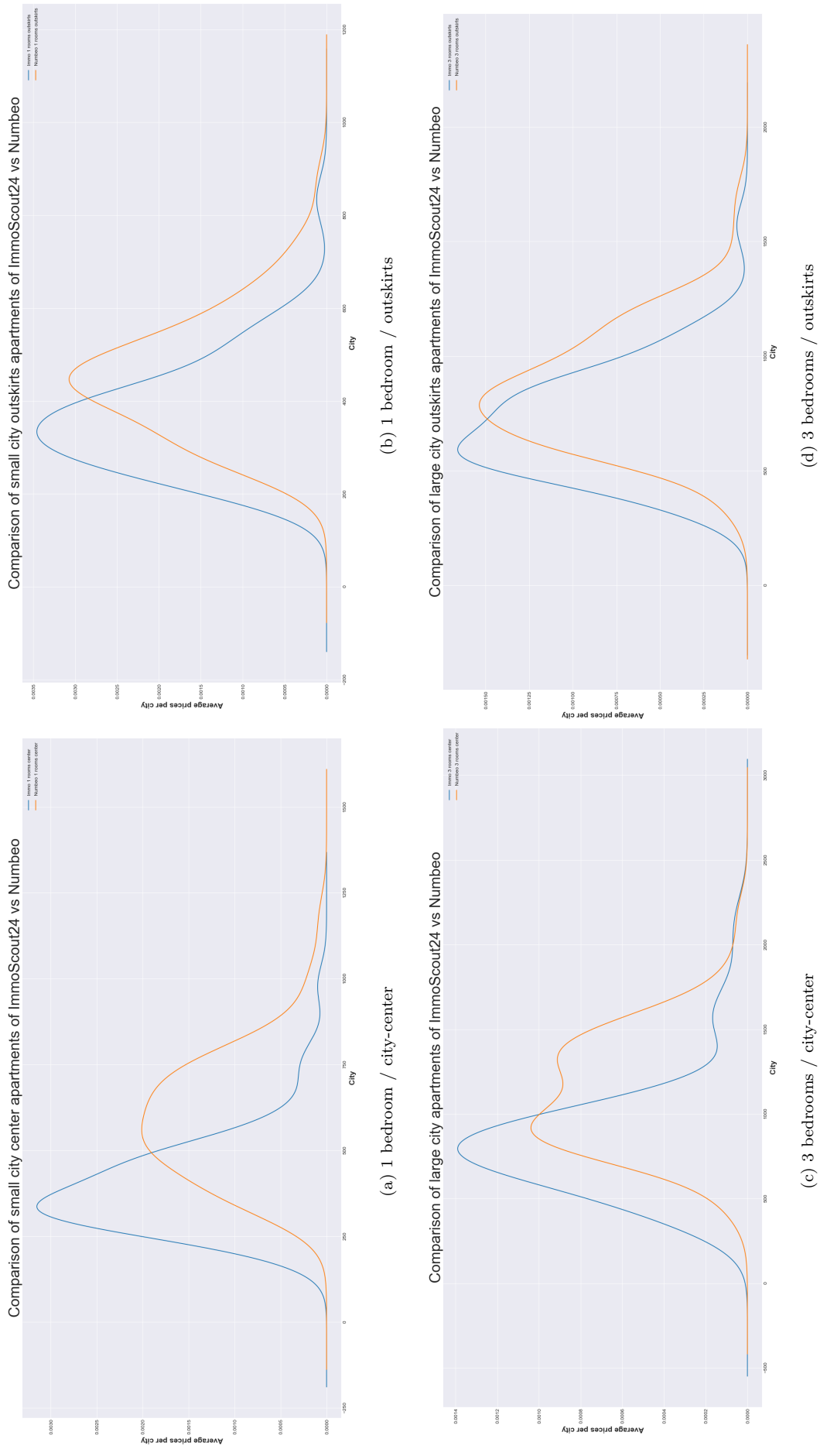
38 A continuous probability distribution describes the probabilities of occurrence for a continuous random variable that is differentiable at any point x. Its counterpart would be a discrete probability distribution.

39 *Non-parametric* methods refer to the branch of statistics that is not-based on pre-defined probability families and their assumptions, but often lacks the generalizability of the commonly encountered statistical patterns.

40 *Supremum* is a concept that is intuitively closely related to the maximum-definition. Mathematically, it is the *least upper bound* of a set, if compared to another set, so the least element of a Set T that is at least greater than or equal to all elements in Subset S.

41 *Almost sure convergence* implies convergence in probability (by Fatou's lemma), and hence implies convergence in distribution. It is the notion of convergence used in the strong law of large numbers.

42 The most commonly used *Pearson correlation coefficient* suffices to build intuition. It represents a measure of linear relationship between two variables and is bounded between -1 (perfect inverse dependence) and +1 (perfect dependence).



(a) 1 bedroom / city-center

(b) 1 bedroom / outskirts

(c) 3 bedrooms / city-center

(d) 3 bedrooms / outskirts

Figure 3: Kernel density comparison of the underlying distributions of the data. Certain discrepancies seem to persists, but general tendencies are captured well betweenn the two fully unreated sources.

	city-center	**outskirts**
1 bedroom	0.8131	0.7403
3 bedrooms	0.7561	0.8148

Table 1: A comparison of the listed correlation coefficients hints at a significant relationship between the scraped data of ImmoScout24 and Numbeo.

set). Highest density points are impressively close to one another, though, for the fully disconnected data sources and therefore hint at a data reliability, which allows for the subsequent analysis of the upcoming sections.

4.2 Kolmogorov-Smirnov test

The KS-test results are fourfold for the four different possible combinations between 1-bedroom/3-bedrooms and city-center/outskirts.

Table 2 compares these four distributions between ImmoScout24 and Numbeo and highlights their significance and respective p-value.

	city-center	**outskirts**
1 bedroom	0.582 ($7.611 \times e^{-11}$)	0.402 ($2.999 \times e^{-5}$)
3 bedrooms	0.447 ($2.009 \times e^{-6}$)	0.296 ($5 \times e^{-5}$)

Table 2: Statistical results of the KS-test to compare the 4 target distributions show that all of them are significantly different. Low p-values indicate significance of the test statistic to falsify the null-hypothesis that the two samples stem from the same distributions. Result presentation: "Test-statistic (p-value)"

This result hints at an inherent difference in the two datasets sampled. This discrepancy can stem from different reasons, which Section 5 discusses in detail in, but might either indicate a bias of the data entered into Numbeo, a bias of the rent offers on ImmoScout24 or a potential seasonality effect (actually the distributions would be the same, but scraping the data for six weeks might not reflect the same extended time-series data as Numbeo).

Figure 4 serves to explain better the reasoning, why the ImmoScout24 rent data still offers a valid input to answer RQ2 and RQ3. As we can see, the distributions are decently similar, but rent figures from Numbeo seem to be consistently higher. It is essential to notice that the order between the different accounts differs but for improved readability, the difference determines the ordering of the cities. Therefore, one might lose some comparability between the individual subplots but gains an improved sense of how the actual relationship of the two distributions behaves, which for data validity reasons is more interesting. This overestimation of Numbeo figures, which contains manually entered on their website, hints at the effect that people, who are unhappy with their rent prices (or for Numbeo, with their purchasing power in a given location n general), are more likely to use this outlet to let other consumers know (see for example Krstić Vukelja and Runje [20]). This subsection closes the results discussion for RQ1, in which we asked for data validity and potential overestimation of the rent prices. We were able to validate

our hypothesis (see section 3) that the subjective perception of rent prices is not aligned with the actually observed prices listed online.

4.3 Web-scraping

This section details the results from pure data scraping. Besides serving as a reference point for individual city rent price lookups, one of the main purposes of the tables in this section is the illustration that a reader cannot understand the data well without the context of visualization. Section 4.4 attempts to close this gap. Important to notice is that the short rent price overview in Table 3 only showcases 10 of the 14 cities used in the interactive map. This discrepancy is because the large dataset off Numbeo, even though in total including 67 cities, only covered 10 of the 14 largest areas in Germany. Therefore, Table 3 would have lost some educational value by including all 14 cities. The Section A of the Appendix displays the long table with the scraped and the listed rent prices for all 67 cities but is due to space limitations not included in the main body of this research work. Table 5 presents almost the same table, without the information off Numbeo, but extended to all 14 cities.

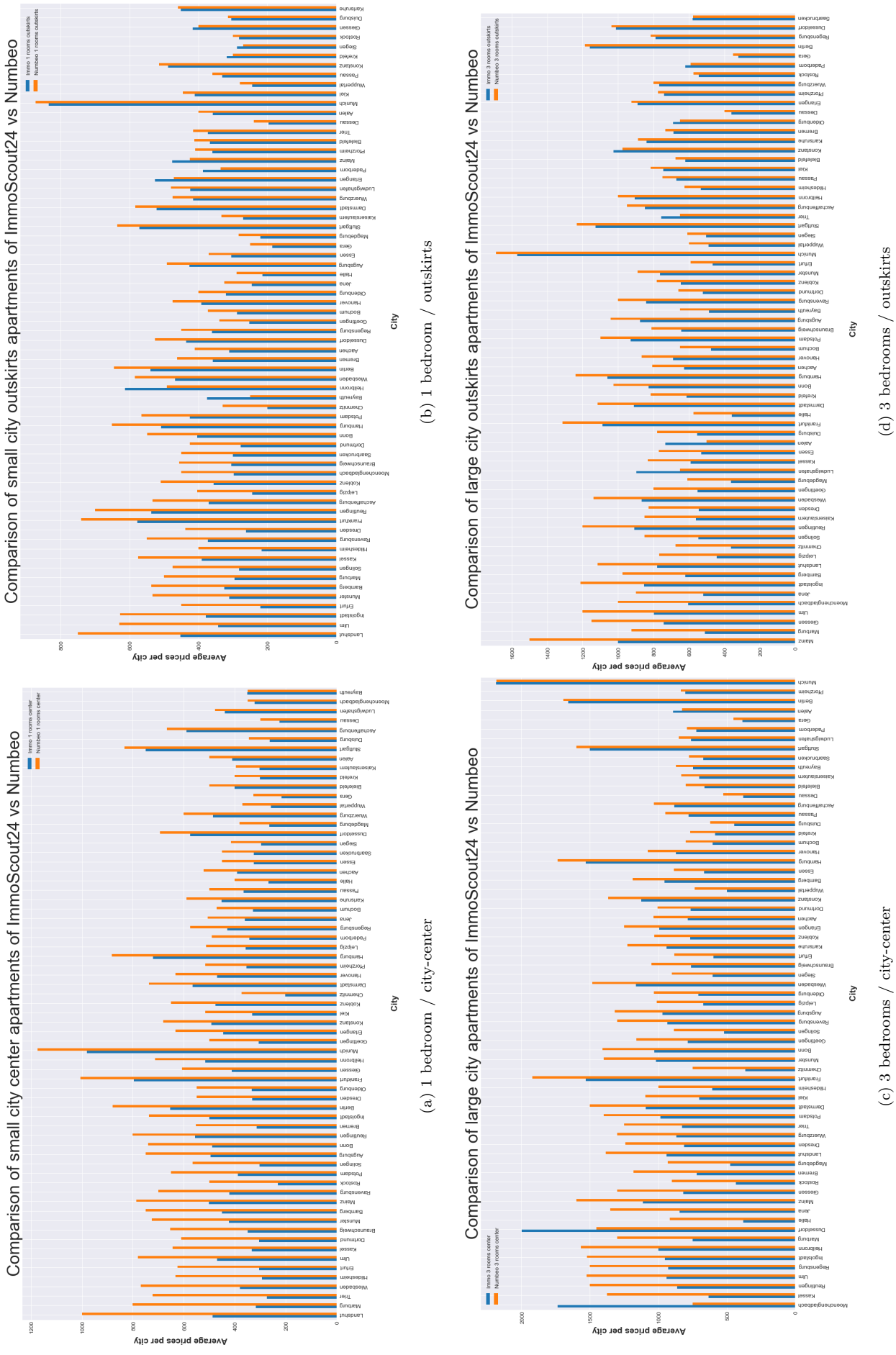
	City-center (<5km)				**Outskirts (>5km; <15km)**			
	1-room		**3-rooms**		**1-room**		**3-rooms**	
City	**IS***	**NB****	**IS**	**NB**	**IS**	**NB**	**IS**	**NB**
Berlin	653 €	880 €	1,658 €	1,693 €	539 €	645 €	1,159 €	1,187€
Bremen	314 €	553 €	719 €	1,181 €	359 €	462 €	686 €	733 €
Dortmund	304 €	610 €	762 €	1,005 €	277 €	425 €	522 €	660 €
Duesseldorf	575 €	694 €	1,999 €	1,452 €	436 €	527 €	1,011 €	1,037 €
Essen	325 €	450 €	664 €	885 €	306 €	370 €	529 €	771 €
Frankfurt	797 €	1,006 €	1,530 €	1,921 €	578 €	741 €	1,089 €	1,314 €
Hamburg	721 €	882 €	1,529 €	1,736 €	509 €	652 €	1,060 €	1,241 €
Hanover	469 €	633 €	872 €	1,076 €	392 €	475 €	689 €	868 €
Munich	980 €	1,174 €	2,186 €	2,182 €	835 €	873 €	1,570 €	1,690 €
Stuttgart	749 €	833 €	1,500 €	1,599 €	572 €	636 €	1,127 €	1,233 €

Table 3: Overview of the table comparing the rent prices for the small subset of 10 (only for 10 out of the 14 cities in subsequent analyses, the Numbeo data was available at the time of data retrieval) cities to illustrate the data verification done for the larger set, shown in Table 5. ImmoScout24 data for 1-3 bedroom apartments (inside vs outside center) is compared to the equivalent data off Numbeo.

** IS: ImmoScout24*

*** NB: Numbeo*

Table 4 serves as the only missing ingredient to continue towards the main contribution of this research endeavor, namely the economic data that seeks to explain the rent price differences on the map. The column includes infor-



(a) 1 bedroom / city-center

(b) 1 bedroom / outskirts

(c) 3 bedrooms / city-center

(d) 3 bedrooms / outskirts

Figure 4: Comparison for the actual average prices between ImmoScout24 and Numbeo rent prices. The data shows that the general tendency is captured well, but that Numbeo prices are almost entirely reported higher than the actual prices scraped off ImmoScout24.

mation about population as a starting point for overall size, to purchasing power as an intuitive explanation for the cost of living and, therefore, rent in a given city. It further introduces figures for unemployment, crime rate and as a measure of proactivity by the different cities, a rate of new apartments built. It is crucial to notice that not all measures are available on the city level, but on state-level (*Bundesland*), which leads to the many identical values for closely located cities in the list.

	Economic indicators				
City	**Population**	**Purchasing power index**	**Unemployment rate**	**Crime rate**	**New apartment rate**
Berlin	3,613,495	91.5	7.7%	14%	1.28%
Bremen	568,006	93.2	9.8%	11%	0.82%
Dortmund	586,600	91.7	6.4%	11%	0.58%
Duesseldorf	617,280	117.7	6.4%	10%	0.58%
Dresden	551,072	91.1	5.2%	9%	0.60%
Essen	583,393	97.2	6.4%	9%	0.58%
Frankfurt	746,878	114.2	4.3%	16%	0.88%
Hamburg	1,830,584	109.8	6.0%	12%	1.31%
Hanover	532,163	97.9	5.8%	13%	0.74%
Koeln	1,080,394	106.5	6.4%	12%	0.58%
Leipzig	581,980	87.5	5.2%	12.%	0.60%
Munich	1,456,039	134.8	2.8%	7%	1.17%
Nuernberg	515,201	103.9	2.8%	8%	1.17%
Stuttgart	632,743	112.5	3.2%	9%	0.84%

Table 4: Summary of the main economic data. The table displays, besides the rent date in table 5, the main input data for the visualization in this section.

Table 5 provides an alternate view of Table 3, which on the one hand, extends the list of included cities of 10 to all 14 in the visualization, but lacks the Numbeo information since Numbeo misses 4 of 14 cities.

4.4 Visualization

The main contribution of this research paper is the visualization[43], which tries to allow for an intuitive and play-like approach to complicated and intricate interactions of rent prices with multiple dimensions of potential drivers. Since the predominant value of the visualization lies with its interactive nature, the illustration in this write-up limits itself

43 This footnote serves as a quick explanation to why no correlation coefficients between rent prices and economic indicators are offered: (1) The sample size will not allow for interpretative results - with only 14 cities and often even less economic indicators (sometimes just one per state), the quantitative analysis might rather confound than facilitate interpretation, which stands in contrast to the overall research endeavor. (2) Since the main contribution is to *not* focus on single effects, but to let difficult interactions show the intrinsically complicated nature of rent price interpretation, the correlation coefficients, or even linear regression models with interactions, would need to be expanded to all possible combinations of economic indicators. This result demonstration would suffer from point (1) and will damage the playful approach I hope to achieve through an interactive visualization.

	Average rent prices of *ImmoScout24*			
	1-room apartments		3-rooms apartments	
City	**City center (<5km)**	**Outskirts (>5km; <15km)**	**City center (<5km)**	**Outskirts (>5km; <15km)**
Berlin	635 €	500 €	2,129 €	1,052 €
Bremen	312 €	347 €	680 €	670 €
Dortmund	308 €	279 €	777 €	484 €
Duesseldorf	662 €	409 €	2,627 €	813 €
Dresden	336 €	264 €	790 €	533 €
Essen	336 €	293 €	653 €	507 €
Frankfurt	772 €	570 €	1,510 €	1,071 €
Hamburg	739 €	478 €	1,405 €	1,012 €
Hanover	446 €	358 €	839 €	670 €
Koeln	652 €	495 €	1,227 €	935 €
Leipzig	369 €	245 €	684 €	441 €
Munich	937 €	797 €	2,108 €	1,529 €
Nuernberg	536 €	577 €	906 €	872 €
Stuttgart	742 €	575 €	1,470 €	1,114 €

Table 5: Rent price overview for the scraped cities. This data is the average of the over 20k single rent offers scraped from ImmoScout24. The prices might in this table might not reflect the very latest results scraped, since the scraping efforts are continuously conducted, to best alleviate seasonality-effects.

to exemplary images, but the reader is strongly encouraged to download the .zip file '*map_folder.zip*' here[44]. The password to download the file is *CSSSeminar2020*. After unzipping[45], the main file in this folder is the *map.html* file, please leave the actual file (*map.html*) and the necessary folder *map_files* always in the same folder. [46]

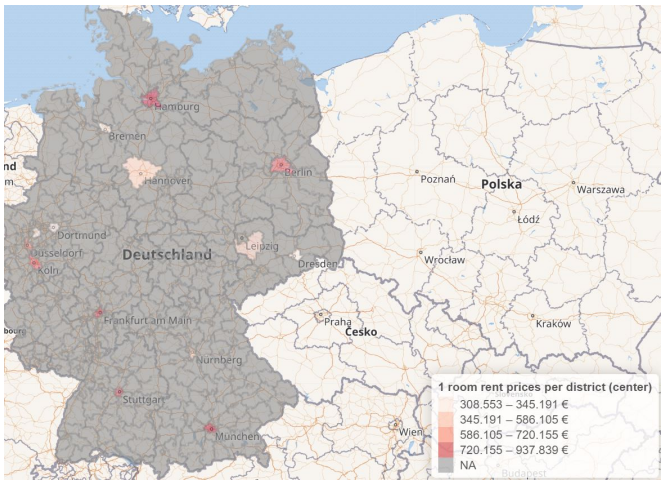
Figure 5 illustrates the rent information that is available on the left tab by clicking on the respective radio button. Here the hovering functionality to show the individual rent prices for each one of the 14 cities is ignored but adds in the .html version for readability and intuition. As one can see, the general tendency for a given city to be either high-priced or cheap does not differ substantially between the categories. Munich places in a dark red for all four subfigures, while Hanover builds the lower spectrum with a light beige tone. This visualization demonstrates how a convoluted table of dense information, such as Table 5, becomes alive in a simple spatial illustration.

This demonstration only addresses one part of RQ2, namely how far rent prices, and bubbles, are predictable by economic indicators - and consequently, if cities prove to be overpriced or if the higher cost is just a premium the consumer pays for a higher standard of living.

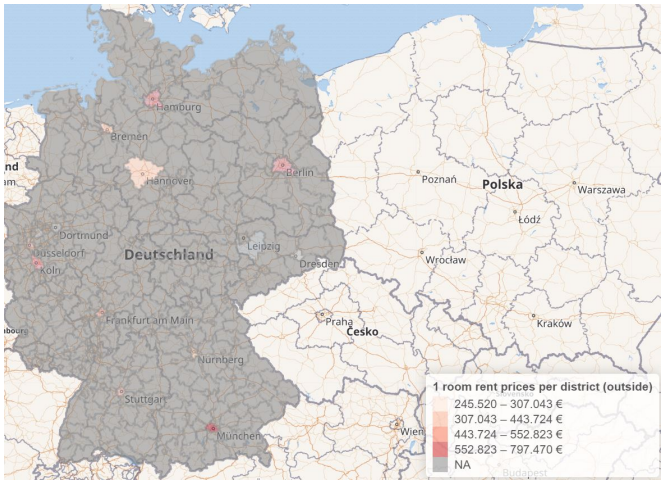
44 In case the hyper reference does not work, please find the file under https://cloud.uni-konstanz.de/index.php/s/ad2MMPbzG4sP6cy

45 Common programs to unzip a file are *WinZip* or *WinRar*. https://support.microsoft.com/en-us/help/14200/windows-compress-uncompress-zip-files provides a quick (windows-based) tutorial.

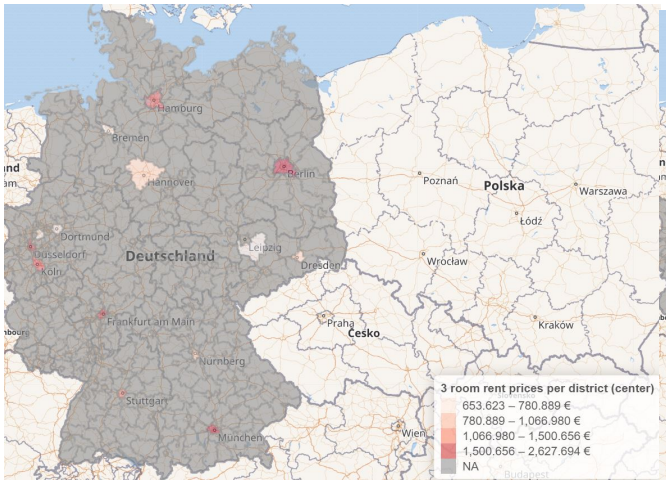
46 It is vital to notice that details of rent prices between the tables provided (see Tables 6 and 5) and the HTML file differ since distinct scraping efforts acquired the data for the large group of cities and the smaller subset. The differences are minuscule and will not compromise interpretability, but might lead to confusion for the interested reader, who compares the numbers in detail.



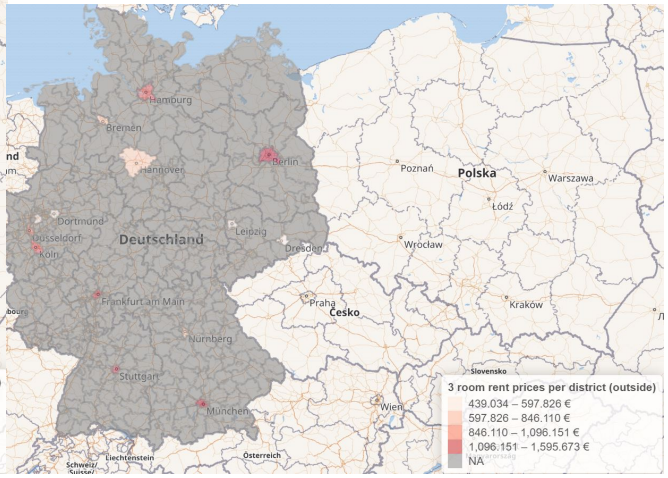
(a) 1 bedroom / city-center



(b) 1 bedroom / outskirts



(c) 3 bedrooms / city-center



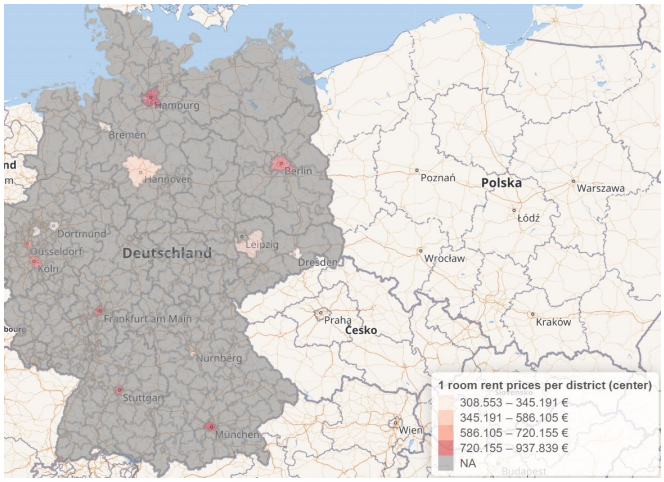
(d) 3 bedrooms / outskirts

Figure 5: Comparison of rent prices for the different categories of the interactive map shows similar tendencies for 1 and 3 bedroom apartments and either city-center or outside of city.

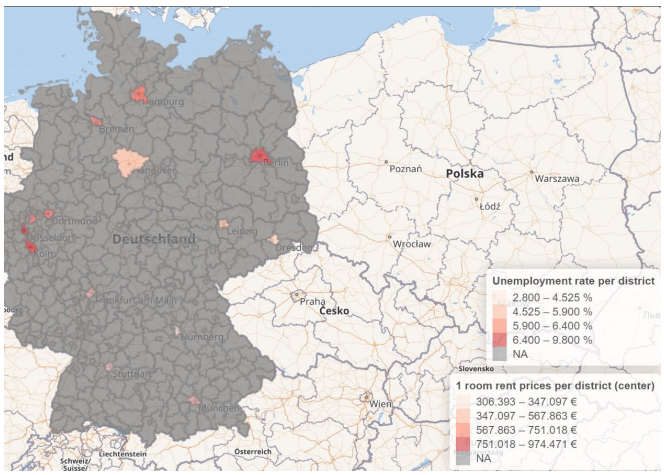
For the remainder of this result discussion, the discussion discourses one room rent prices for city-center apartments. The complexity of the interactions bought through visualization can only very superficially explain interactions on paper by referencing every single scenario. The reader is highly encouraged to follow the instructions in Figure 6 to build intuition with the interactive map itself.

Different interesting effects stick out to the reader. The top right Figure 6a is identical to the top right figure in Figure 5, namely Figure 5a. It generally only depicts the renting landscape in Germany and shows that certain regions lean towards the top end of the range, such as Munich, Hamburg, Stuttgart, or Frankfurt, while other cities appear to be cheaper, such as Leipzig, Hanover, and Dortmund.

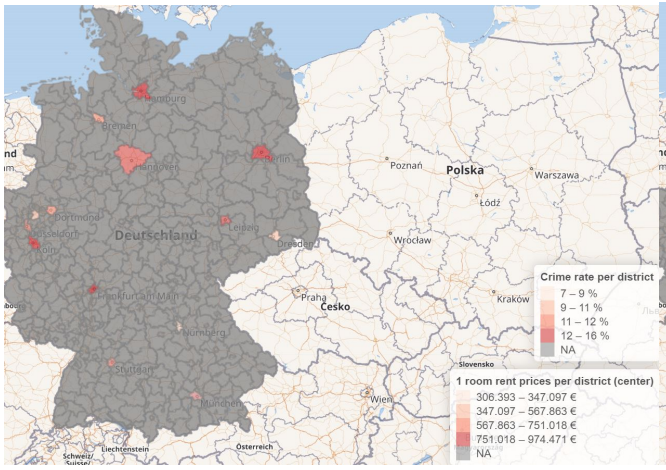
The second subfigure on the top right, Figure 6b. combines the one-bedroom, city-center layer with a corresponding layer for unemployment rates. Since both layers refer to the same color-scale, a very high unemployment rate (dark red) would worsen the visual effect of high rent. The opposite is also exact: while a rent paints the region in dark red, a low employment rate would decrease the effect and the result would appear more beige than



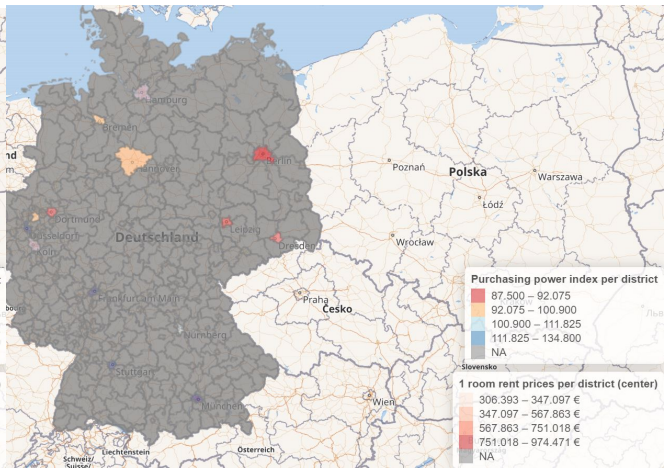
(a) 1 bedroom / city-center alone



(b) 1 bedroom / city-center with unemployment



(c) 1 bedroom / city-center with crime-rate



(d) 1 bedroom / city-center with purchasing power

Figure 6: Illustration of the power of the interactive visualization to combine different effects that might drive rent prices to allow for better intuition, which regions are substantially overprices and for which cities inhabitants might just pay a premium for higher living standards.

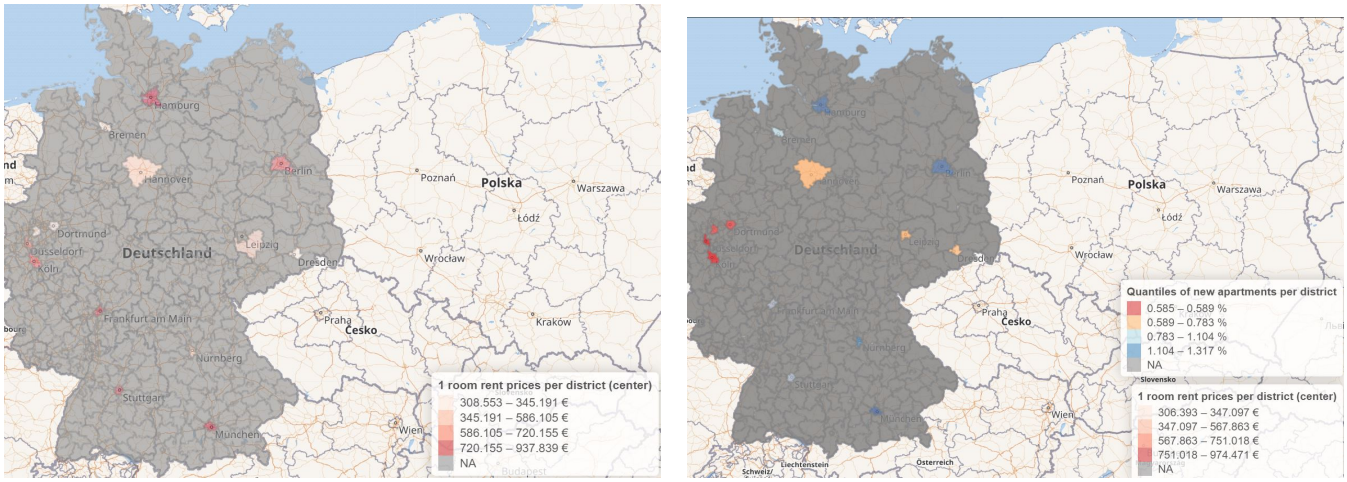
the baseline image to the top left. The actual results illustrate certain exciting features, such as that Cologne and Dusseldorf, which portray average rent prices, are dark red. This dyeing illustrates that these regions pair average rent prices with a very high unemployment rate for the state of *North-Rhine Westphalia*, which in combination appears to be a worse combination than the isolated rent price info. Munich, a notoriously pricey candidate for the German rent market, has almost lost all its color. Bavaria experiences almost the lowest unemployment rates for all of Germany, so this picture might solve the puzzle, what Munich's inhabitants are receiving as consumers. Other honorable mentions are Berlin, which appears to be a worse pick in terms of rent + unemployment rate, as well as Bremen, which appears decently cheap, but with extraordinarily high unemployment rates.

The third subfigure (Figure 6c) shows the same pairing, but with the state-wide crime rate. Berlin and Hamburg stand out as hubs of high crime-rates, paired with their rent information. The *Ruhr Area*, around Dortmund and Essen, suffers from high crime rates in *North-Rhine Westphalia*, but their below-average rent prices offer an about average or still below average combined effect. A notable mention is Hanover, which stood out as a haven of low

rent so far, but which is also tainted in red now, due to its high crime rates.

While the last two indicators give an insight into the quality of living, the last subfigure, Figure 6d), provides more essential insights into what the rent prices translate to in terms of real purchasing power. The results are astonishing. Munich and Stuttgart, leaders in terms of high rent prices, are suddenly appearing in dark blue[47], due to their unproportionally high purchasing power. Cologne and Hamburg disappear off the radar of unlivable cities as well, while Dortmund appears to be a worse fit, namely: even if rent is low in Dortmund in absolute terms, the money one generally makes, covers rent expenses worse than in comparable other cities. The entire East German fraction paints in dark red, which underlines the low purchasing power, paired with comparatively high rent prices in the region (see for a discussion on the lack of catching up of East German states and East European cities in general in terms of purchasing power and economic activity Wießner [36].). This example closes the discussion of the results for RQ2.

RQ3, which mainly addresses the question, if city officials try to counteract the rising shortage of houses and rent prices (supply theory of housing), can also be intuitively answered via the visualization. Additionally to the economic indicators, another layer, which adds the rate of new houses built in a city as a counterweight, and possibly as a color layer, which demonstrates new tendencies. The results for our toy example are highlighted in Figure 7.



(a) 1 bedroom / city-center alone

(b) 1 bedroom / city-center with new apartment rate built rate

Figure 7: Comparison of rental prices for 1 bedroom in city center to the left and the same image overlaid with the new apartment rate for the given cities shows that certain cities very proactively address the high rent prices.

The strong tendencies become immediately evident through the intense color scheme. While Munich ends up dead-last in terms of rent prices, it seems as if the city is aware of the problem and attempts to alleviate the problem by rising newbuilding rates to the top of the sampled cities. Berlin and Hamburg, even if not as hard-hit by the rent prices, address the issue. Cities in the *Ruhr Area* end up much worse in comparison. The rent prices do not spike as high as in the South, but the new building rates are at the bottom of the pack, which might indicate future problems. It is important to notice that the rate of new apartment buildings built in a city is a relative value

47 The blue color represents here the two-tailed characteristic of the metric of *purchasing power index*. Its stronger scheme can better overlay the rent price colors and, therefore, creates a more explanatory effect for the visualization.

based off the existing apartments, which makes this result even more astonishing for large cities, such as Berlin and Hamburg since 1.3 % building rate of a large already existing body of apartments translates to a high number of new houses. This one-dimensional comparison does not explain much since all cities deal with different federal and state budgets and various geographical limitations (the *Ruhr Area* is highly populated, which might limit the extent to where city-officials can plan new apartments), but it does allow for a new perspective on the often over-simplified problem of rental peaks.

The effects of multiple indicators are available as well, but for the sake of readability is left out of the result discussion of this paper. Potentially exciting combinations to try is a particular rental category vs. unemployment and crime rate together, or purchasing power and new apartment rate together. These combinations allow for an interpretation along the lines of "is it worth it to live in a city with bad living conditions, i.e., high unemployment and high crime rate?", or "Is it costly to live in X, if our money can buy is more and rent prices might drop due to many new buildings built?". The visualization, which was always just intended as a starting point, might then quickly reach its limits since the color intensity of the individual layers might convolve the distinct effects.

5 Possible extensions and discussion

This section will discuss shortcomings, challenges for future work to overcome, and underlying assumptions that allow this paper's main findings.

5.1 Challenges

Since the scope of this endeavor addresses comprehensive questions relating to mass human behavior and its intersection with external factors, simplification for the sake of interpretation comes naturally. It is almost philosophical to ask how much can be simplified before the research project ends up as a thought experiment without any external validity. Some of the potentially critical, assumptions that are inherent to the paper's method are:

1. The data lists off Numbeo and ImmoScout24 are comparable.
2. The small time window of data retrieval does not suffer from seasonality effects.
3. For the economic factors, which are only available on state-level, the generalization to city-level holds.
4. The factors are exegeneous[48], they are not responses to or at least impacted by the actual variable in question, i.e., the rent prices (assume no endogeneity[49] problems).
5. The rent prices are representative for the population. No selection bias is present for ImmoScout24 listings.

48 *Exogeneity* describes a factor that is not impacted by the system, but represents an external factor to it. F. ex. the weather might change wheat prices, but wheat prices do not impact the weather. This is often in economics also referred to as an *external shock*.

49 *Endogeneity* is the counterpart to *Exogeneity*, describing a factor that interacts with the variables of a system. This characteristic can become troublesome, if the model assumes factor-independence.

Since these assumptions are cherry-picked to represent the polarizing subset of critical assumptions, which might compromise validity of the research findings, alleviating actions towards these challenges are listed below and will address all the individual problems respectively. Important to notice is that some of these complications are fundamental to the simplifications necessary for abstraction, therefore full external validity will be impossible to achieve[50].

1R. The idea to scrape enough data to allow for a valid comparison is disqualified according to section 4.2, which states that the samples do not originate from the same underlying distribution. Does that disqualify ImmoScout24 as a data source for our analysis purposes? We have explored the distributions in Figure 4, which allows us to informally distinguish the samples to behave similar, but with generally a larger value for the Numbeo figures. This discrepancy is due to multiple possible interpretations, such as an overestimation of the numbers, if entered manually, a selection bias for people on Numbeo (unhappy consumers might be more likely to share their experience), or a selection bias on ImmoScout24 (ImmoScout24 might qualify as a low-quality apartment market place, and other sources might better reflect actual rent prices.). While all these objections might be valid, ImmoScout24 represents one of the most prominent rental markets online in Germany and as long as the respective order of prices remains intact, the interpretability of the analysis does not suffer from a parallel shift of prices from ImmoScout24 to Numbeo.

2R. Seasonality and selection bias is an omnipresent complication within most empirical studies. Since we do not deal with classical time-series data, techniques to eliminate fixed effects[51] are not at free disposal. Continuation of the web-scraping efforts beyond the end of this study allows for a better estimation of the effect, but even through a potentially biased dataset, general interpretations and first insights will be possible. To the author's knowledge, no specific seasonality taints the data quality of rent prices for November and December any more than for other seasons. See section 5 for additional approaches to address concerns regarding the sampling bias of the data.

3R. Generalization of economic factors are complicated. Since Germany is geographically relatively small, the factors, which are broken-down from state- to city-level, will closely align with their correct figures. Often the numbers listed are estimates themselves derived from the main economic hubs in the state. For certain cities, the state estimate is congruent to the city-estimate, due to political affiliation, see for example Berlin, Bremen, and Hamburg.

4R. This problem is usually the most complicated to address. It bridges the gap between establishing causality from correlation and often represents the hurdle that most empirical research papers struggle with. While an improved data source usually addresses data quality concerns, the research methodology in general is a major limitation without easy fix. How much are unemployment and crime rate driven by bizarre rent prices? Are new building rates of new houses a consequence of a ballooning political budget, or consequential to the rent

50 X**R**. refers to the respective response to the numbered challenge above.

51 *Fixed Effects* refer to stable effects, such as location of cities. While in single observations, these factors might convolute the analysis, for subsequent samples, one can subtract the model specifications to eliminate constant terms.

price epidemic? The questions to this paper's interpretations are obvious, but might fail to recognize the main contribution of this it. To establish a starting point, to hint in the right direction, and to highlight some surprising correlations in a less rigorous and more descriptive way, does not avoid the author's responsibility to ask complicated statistical methodological questions, but allows the paper to speak openly, without excessive filtering. How much of these findings are spurious phenomena is up for exploration in extensions to this study.

5R. The data might very well experience a certain bias. A possible alleviations is to merge in additional data sources, such as student apartments[52], or high income apartments[53]. Generally, a certain downward-bias will persist, since high-income individuals are more commonly acquiring real-estate through a private agent, or personal connections. For the given research question of representing the income of the common population, the tails of the distribution might not add much explanatory value, though.

5.2 Extensions

The possible extension to a descriptive study are numerous, therefore this section only hints at the most obvious augmentations that come to mind:

- Prolonging the research time-frame. By extending the data collection period, the data establishes more profound results for the discrepancy between Numbeo and ImmoScout24 sampling, avoidance of the seasonality effect of the rent prices, as well as more robust results for the economic indicators (if extended and tested for different years).
- Robustness checks for an extended time period. This follow-up improvement, allows for sampling of sub-periods, which then verify certain correlations, or label them as seasonal on specific months, seasons, or years.
- Include more than the 14 largest metropolitan areas. This augmentation of the dataset allows for a more insightful visualization, numerous robustness checks on city level (does this effect only hold for large metropolitan areas, why?), and the alleviation of a main critique addressing the generalization of economic factors.
- Include additional factors to allow for other effects to explain rent prices. Educational figures (as mentioned in section 2.5), more distinct crime-rate effects (stealing vs. murder, etc.), or infrastructural indices might give a more coherent picture of what particular price premia include.
- Follow-up studies might approach this study from a more statistical standpoint, which overcomes the barriers of a descriptive study, but will need to introduce methods for more causal relationships, based on sound theory and a rich dataset.

52 https://www.wg-gesucht.de/
53 https://www.luxuryestate.com/germany

6 Conclusion

This study tries to identify a first starting point to the question: *"Are we currently experiencing a rent price bubble, and what are its main drivers?"*. To introduce this ambitious research endeavor, the paper structures the approach along with three main research questions with accompanying theses. The general logic follows the pattern of establishing data validity, finding causes of high and low rent prices and building an intuition if city-officials are aware of the problem.

To acknowledge these questions, the paper has settled for an interactive approach with the reader by building a map visualization of the rental prices in Germany and trying to mimic interactions by overlaying the coloring of the largest 14 metropolitan areas between economic factors and rent prices. This concept allows for playful cooperation with the interested reader, of which the paper illustrates exemplary some of the potential insights of the many combinations. Some highlights mentioned in section 4 are that rental prices seem to be costly for the subset of commonly expected expensive cities but that a high purchasing power often offsets this effect, which means that in relative terms, a city, such as Dresden might become more expensive to live in than Munich. It also seems apparent that inhabitants of a city pay for certain perks, such as a low unemployment rate or low crime rates in high-cost cities, such as Stuttgart and Nuremberg, which demonstrates that Cologne might be overpriced.

The paper discusses potential shortcomings, alleviating actions, and possible extensions addressing the challenges or widening the research scope in detail. The data quality and short retrieval period, as well as potential endogeneity concerns, disqualify the establishment of causal relationships but promote further work in the field, which can benefit from a systematic and descriptive study that continually improves its quality through ongoing data scraping efforts.

A Appendix and supplementary material

	City-center (<5km)				Outskirts (>5km; <15 km)			
	1-room		3-rooms		1-room		3-rooms	
City	**IS***	**NB****	**IS**	**NB**	**IS**	**NB**	**IS**	**NB**
Aachen	391 €	522 €	785 €	1,035 €	311 €	411 €	627 €	807 €
Aalen	410 €	500 €	891 €	825 €	359 €	400 €	733 €	500 €
Aschaffenburg	590 €	666 €	883 €	1,033 €	370 €	533 €	849 €	950 €
Augsburg	496 €	749 €	970 €	1,318 €	426 €	492 €	877 €	1043 €
Bamberg	451 €	750 €	956 €	1,187 €	325 €	537 €	620 €	975 €
Bayreuth	352 €	350 €	746 €	870 €	375 €	250 €	487 €	650 €
Berlin	653 €	880 €	1,658 €	1,693 €	539 €	645 €	1,159 €	1,187 €
Bielefeld	401 €	500 €	663 €	800 €	367 €	412 €	620 €	675 €
Bochum	329 €	471 €	602 €	800 €	289 €	373 €	475 €	650 €
Bonn	489 €	740 €	1,028 €	1,409 €	404 €	549 €	828 €	1,027 €
Braunschweig	349 €	654 €	762 €	1,050 €	305 €	456 €	643 €	811 €
Bremen	314 €	553 €	719 €	1,181 €	359 €	462 €	686 €	733 €
Chemnitz	201 €	374 €	364 €	750 €	200 €	330 €	362 €	675 €
Darmstadt	565 €	736 €	1,093 €	1,500 €	522 €	583 €	911 €	1,117 €
Dessau	225 €	300 €	380 €	525 €	197 €	240 €	360 €	400 €
Dortmund	304 €	610 €	762 €	1,005 €	277 €	425 €	522 €	660 €
Dresden	332 €	550 €	811 €	1,242 €	262 €	438 €	543 €	828 €
Duisburg	263 €	345 €	446 €	620 €	305 €	315 €	553 €	780 €
Dusseldorf	575 €	694 €	1,999 €	1,452 €	436 €	527 €	1,011 €	1,037 €
Erfurt	304 €	625 €	595 €	883 €	220 €	450 €	466 €	590 €
Erlangen	445 €	633 €	993 €	1,250 €	526 €	472 €	890 €	925 €
Essen	325 €	450 €	664 €	885 €	306 €	370 €	529 €	771 €
Frankfurt	797 €	1,006 €	1,530 €	1,921 €	578 €	741 €	1,089 €	1,314 €
Gera	216 €	327 €	384 €	450 €	186 €	250 €	321 €	350 €
Giessen	411 €	607 €	819 €	1,300 €	417 €	400 €	742 €	1,150 €
Goettingen	307 €	500 €	786 €	1,160 €	252 €	340 €	551 €	800 €
Halle	268 €	400 €	379 €	916 €	215 €	290 €	359 €	575 €
Hamburg	721 €	882 €	1529 €	1,736 €	509 €	652 €	1,060 €	1,241 €
Hanover	469 €	633 €	872 €	1,076 €	392 €	475 €	689 €	868 €
Heilbronn	516 €	712 €	1,000 €	1,566 €	614 €	492 €	907 €	1,000 €

Table 6 continued from previous page

	City-center (<5km)				Outskirts (>5km; <15 km)			
	1-room		3-rooms		1-room		3-rooms	
City	IS*	NB**	IS	NB	IS	NB	IS	NB
Hildesheim	293 €	633 €	606 €	1,000 €	217 €	400 €	534 €	625 €
Ingolstadt	499 €	737 €	951 €	1,520 €	379 €	628 €	853 €	1,212 €
Jena	361 €	506 €	843 €	1,350 €	246 €	325 €	519 €	900 €
Kaiserslautern	303 €	395 €	701 €	833 €	270 €	333 €	560 €	850 €
Karlsruhe	451 €	589 €	940 €	1,226 €	451 €	460 €	840 €	888 €
Kassel	333 €	644 €	634 €	1,375 €	391 €	575 €	590 €	833 €
Kiel	331 €	516 €	702 €	1,096 €	411 €	446 €	745 €	816 €
Koblenz	476 €	650 €	765 €	1,030 €	357 €	510 €	645 €	782 €
Konstanz	492 €	680 €	1,125 €	1,366 €	488 €	514 €	1,026 €	975 €
Krefeld	302 €	400 €	586 €	766 €	318 €	300 €	613 €	816 €
Landshut	488 €	1,000 €	939 €	1,383 €	453 €	750 €	778 €	1,116 €
Leipzig	357 €	512 €	672 €	1,010 €	244 €	404 €	443 €	768 €
Ludwigshafen	438 €	478 €	762 €	850 €	424 €	480 €	897 €	650 €
Magdeburg	264 €	381 €	474 €	930 €	220 €	284 €	362 €	610 €
Mainz	501 €	786 €	1,113 €	1,600 €	476 €	425 €	1,001 €	1,500 €
Marburg	318 €	800 €	749 €	1,300 €	295 €	500 €	509 €	925 €
Moenchengladbach	322 €	350 €	1,735 €	750 €	298 €	450 €	604 €	1,000 €
Munich	980 €	1,174 €	2,186 €	2,182 €	835 €	873 €	1,570 €	1,690 €
Munster	423 €	725 €	1,016 €	1,400 €	311 €	534 €	763 €	890 €
Oldenburg	333 €	550 €	706 €	1,033 €	321 €	400 €	690 €	650 €
Paderborn	343 €	491 €	721 €	790 €	388 €	336 €	619 €	590 €
Passau	365 €	500 €	780 €	950 €	331 €	360 €	671 €	750 €
Pforzheim	354 €	516 €	801 €	835 €	360 €	410 €	740 €	775 €
Potsdam	387 €	650 €	984 €	1,400 €	425 €	566 €	929 €	1,100 €
Ravensburg	422 €	700 €	934 €	1,300 €	373 €	550 €	841 €	1,000 €
Regensburg	429 €	575 €	929 €	1,500 €	361 €	450 €	789 €	816 €
Reutlingen	556 €	800 €	863 €	1,500 €	537 €	700 €	909 €	1,200 €
Rostock	230 €	500 €	434 €	900 €	282 €	300 €	544 €	575 €
Saarbrucken	325 €	450 €	671 €	775 €	301 €	450 €	578 €	576 €
Siegen	296 €	415 €	603 €	900 €	288 €	270 €	502 €	610 €
Solingen	303 €	565 €	519 €	885 €	282 €	475 €	548 €	850 €

Table 6 continued from previous page

	City-center (<5km)				Outskirts (>5km; <15 km)			
	1-room		3-rooms		1-room		3-rooms	
City	**IS***	**NB****	**IS**	**NB**	**IS**	**NB**	**IS**	**NB**
Stuttgart	749 €	833 €	1,500 €	1,599 €	572 €	636 €	1,127 €	1,233 €
Trier	275 €	723 €	827 €	1,250 €	373 €	416 €	756 €	650 €
Ulm	469 €	780 €	939 €	1525 €	344 €	630 €	797 €	1,200 €
Wiesbaden	380 €	768 €	1,163 €	1,482 €	468 €	585 €	868 €	1,140 €
Wuerzburg	485 €	600 €	869 €	1,300 €	415 €	475 €	769 €	800 €
Wuppertal	258 €	370 €	499 €	733 €	245 €	280 €	490 €	600 €

Table 6: Overview of the table comparing the rent prices for 67 cities to verify the data quality of subsequent analyses. ImmoScout24 data for 1-3 bedroom apartments (inside vs outside center) is compared to the equivalent data off Numbeo.

** IS: ImmoScout24*

*** NB: Numbeo*

References

[1] Ali Anari and James Kolari. House prices and inflation. *Real Estate Economics*, 30(1):67–84, 2002.

[2] Binyamin Appelbaum. Nobel in economics is awarded to richard thaler. *New York Times*, 2017.

[3] Kelly C Bishop and Alvin D Murphy. Estimating the willingness to pay to avoid violent crime: A dynamic approach. *American Economic Review*, 101(3):625–29, 2011.

[4] Claudio EV Borio and Patrick McGuire. Twin peaks in equity and housing prices? *BIS Quarterly Review, March*, 2004.

[5] William A Branch, Nicolas Petrosky-Nadeau, and Guillaume Rocheteau. Financial frictions, the housing market, and unemployment. *Journal of Economic Theory*, 164:101–135, 2016.

[6] Paolo Buonanno, Daniel Montolio, and Josep Maria Raya-Vílchez. Housing prices and crime perception. *Empirical Economics*, 45(1):305–321, 2013.

[7] Chi-Chur Chao and SH Eden. Capital markets, urban unemployment and land. *Journal of Development Economics*, 38(2):407–413, 1992.

[8] Xi Chen and Michael Funke. Renewed momentum in the german housing market: boom or bubble? *Working Paper*, 2013.

[9] Vincenzo Del Giudice, Pierfrancesco De Paola, and Fabiana Forte. Using genetic algorithms for real estate appraisals. *Buildings*, 7(2):31, 2017.

[10] Robin A Dubin and Allen C Goodman. Valuation of education and crime neighborhood characteristics through hedonic housing prices. *Population and environment*, 5(3):166–181, 1982.

[11] Gabrielle Fack and Julien Grenet. When do better schools raise housing prices? evidence from paris public and private schools. *Journal of public Economics*, 94(1-2):59–77, 2010.

[12] Bernard E Harcourt. Policing la's skid row: Crime and real estate redevelopment in downtown los angeles (an experiment in real time). *U. Chi. Legal F.*, page 325, 2005.

[13] Dirk Helbing. Pluralistic modeling of complex systems. *SSRN*, 2010.

[14] Martin Hoesli. Real estate as a hedge against inflation: learning from the swiss case. *Journal of Property Valuation and Investment*, 12(3):51–59, 1994.

[15] Robert Inklaar and Marcel P Timmer. The relative price of services. *Review of Income and Wealth*, 60(4): 727–746, 2014.

[16] Blair Jenkins. Rent control: Do economists agree? *Econ journal watch*, 6(1), 2009.

[17] Heike Joebges, Sebastian Dullien, and Alejandro Márquez-Velázquez. What causes housing bubbles? a theoretical and empirical inquiry. Technical report, IMK Study, 2015.

[18] Katharina Knoll, Moritz Schularick, and Thomas Steger. No price like home: Global house prices, 1870-2012. *American Economic Review*, 107(2):331–53, 2017.

[19] Michael Koetter and Tigran Poghosyan. Real estate prices and bank stability. *Journal of Banking & Finance*, 34(6):1129–1138, 2010.

[20] Elizabeta Krstić Vukelja and Biserka Runje. Quality service evaluation through the system of complaints and praise. *Interdisciplinary Description of Complex Systems: INDECS*, 12(1):78–91, 2014.

[21] Jacek Łaszek. Housing in consumer's theory. *Working Paper*, 2013.

[22] Stephen Malpezzi and Susan Wachter. The role of speculation in real estate cycles. *Journal of Real Estate Literature*, 13(2):141–164, 2005.

[23] Alex Marsh and Kenneth Gibb. Uncertainty, expectations and behavioural aspects of housing market choices. *Housing, Theory and Society*, 28(3):215–235, 2011.

[24] Andrew Oswald. The housing market and europe's unemployment: a non-technical paper. *Homeownership and the labour Market in Europe*, 1999.

[25] Renata Palen. 5885909b-2ffd-4df3-aab6-40f7b1696ca8. `https://ec.europa.eu/eurostat/documents/2995521/9394392/2-21112018-AP-EN/5885909b-2ffd-4df3-aab6-40f7b1696ca8`, 2018. (Accessed on 12/17/2019).

[26] Narodowy Bank Polski. Rising house prices in europe | | central european financial observer. `https://financialobserver.eu/poland/rising-house-prices-in-europe/`, 0 2019. (Accessed on 12/17/2019).

[27] Matiur Rahman and Muhammad Mustafa. Effects of changes in us unemployment rate, 30-year fixed mortgage rate, and the number of long-term unemployed on changes in us house price index. *Wealth: International Journal of Money, Banking & Finance*, 1(1), 2012.

[28] Jack Rubens, Michael Bond, and James Webb. The inflation-hedging effectiveness of real estate. *Journal of Real Estate Research*, 4(2):45–55, 1989.

[29] Amy Ellen Schwartz, Scott Susin, and Ioan Voicu. Has falling crime driven new york city's real estate boom? *Journal of Housing Research*, pages 101–135, 2003.

[30] Investopedia Staff. Basket of goods definition. `https://bit.ly/2rVQ9mT`, 2019. (Accessed on 12/17/2019).

[31] Alan M Taylor and Mark P Taylor. The purchasing power parity debate. *Journal of economic perspectives*, 18 (4):135–158, 2004.

[32] George E Tita, Tricia L Petras, and Robert T Greenbaum. Crime and residential choice: a neighborhood level analysis of the impact of crime on housing prices. *Journal of quantitative criminology*, 22(4):299, 2006.

[33] Frank Vastmans and Erik Buyst. Interest rates, house prices and the purchasing power for housing. In *ENHR conference paper, Toulouse*, 2011.

[34] Michael Voigtländer. The stability of the german housing market. *Journal of Housing and the Built Environment*, 29(4):583–594, 2014.

[35] Haizhen Wen, Yue Xiao, Eddie CM Hui, and Ling Zhang. Education quality, accessibility, and housing price: Does spatial heterogeneity exist in education capitalization? *Habitat International*, 78:68–82, 2018.

[36] Reinhard Wießner. Urban development in east germany—specific features of urban transformation processes. *GeoJournal*, 49(1):43–51, 1999.