The Big Picture: Assignment 3

Exploratory Data Analysis

Instructors

Dr. Petra Isenberg

Natkamon Tovanich

Authors

Daniel Huf
Mir Wise Khan
Sayor Yusupov

Table of Contents

Introduction

In this assignment, our group performed some exploratory analysis on the Numbeo dataset in order to better understand how structured the data is, and which potential flaws it may contain. In addition, our main topic questions were addressed by creating multiple visualizations on the post-processed dataset.

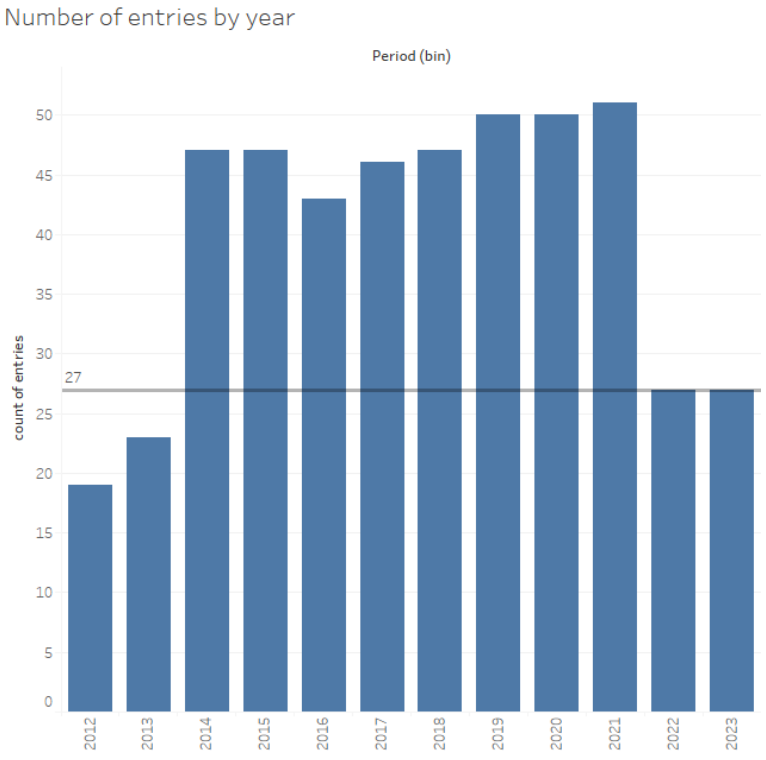
Dataset

Numbeo

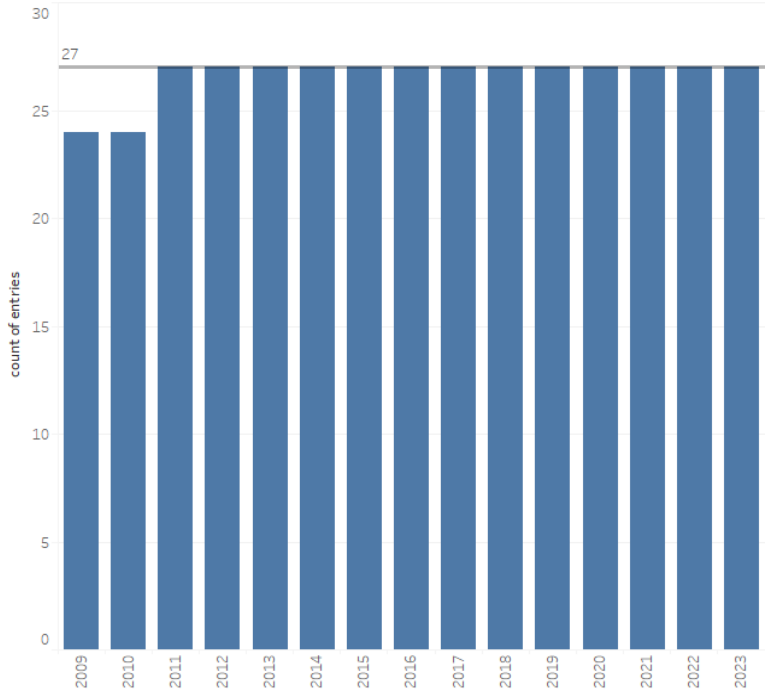
Numbeo is the world's largest cost of living database, providing quality-of-life data in the form of indicators such as housing, crime rates, healthcare, and transport across cities and countries worldwide. Its data is being constantly collected through crowd-sourcing, now containing 9,045,909 prices in 11,524 cities entered by 769,355 contributors.
Through the usage of Numbeo's API, we were able to retrieve valuable insights regarding the cost of living and some life indicators concerning European Union countries. Notably, for this assignment, we focused on the following data:

- **cities**: cities in the database. Omits cities for which there is no data
- **price_items**: items in the main cost of living section
- **country_prices**: current country prices
- **city_cost_estimator**: estimated cost of living for a person or family in a given city
- **historical_country_prices**: historical average prices (per year) in a country
- **country_indices**: Numbeo's indices for a country
- **country_healthcare**: aggregate analysis about health care quality perception in a given country
- **rankings_by_country_historical**: historical rankings by country for a given section at the website (i.e. cost of living, property, crime)

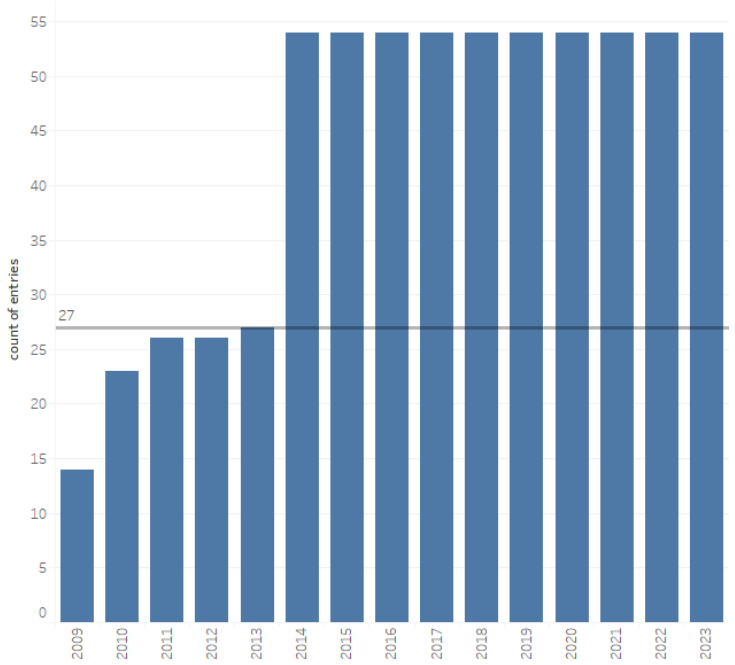
Shape and Structure Analysis



Life Quality Data

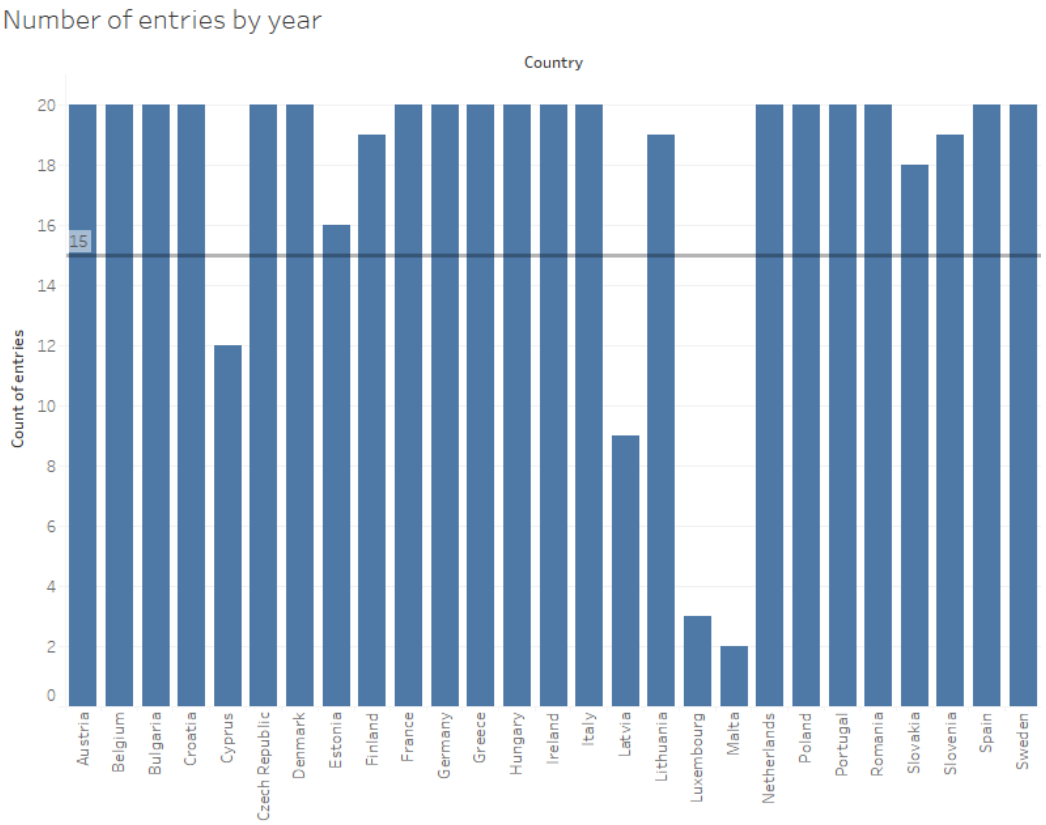


Property Prices Data

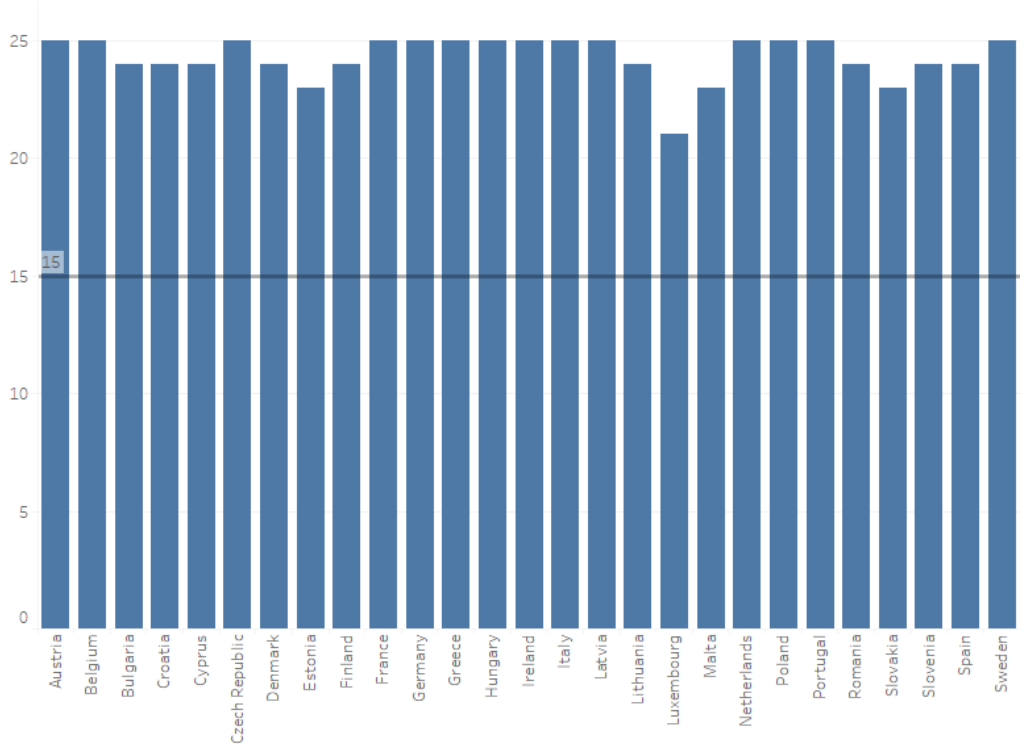


Living Cost Data
Figure 1 (Sayyor Yusupov)

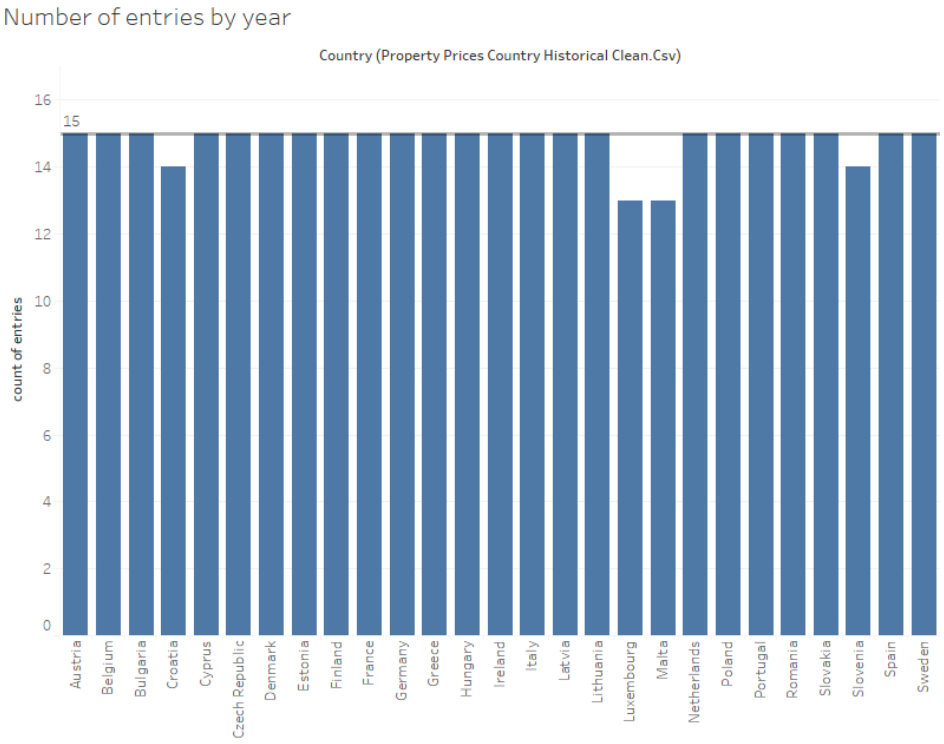
First, we wanted to see if there is any missing/replicated data. So we plotted the number of rows (entries) by year for each data type. There should be one entry for each EU country (27 in total). However, as can be seen from the 3 graphs above, some or all entries for years 2009 to 2013 are missing. Thus, further analysis should be done on data after 2013. Furthermore, some data entries after 2013 have twice the same number of entries.



Life Quality



Living Cost



Property prices
Figure 2 (Sayyor Yusupov)

Now, if we look at the number of rows by country, we can see that some countries (e.g. Malta) have data for only 1 year, while many other countries had more entries than needed. Thus, before the next step, we obtained the mean for multiple entries of a single year and discarded where necessary countries with no entries.

Insightful Discoveries

After having studied the structure of the dataset, it is possible to further the analysis by answering our original research question:

How can fresh graduates strategically select an EU country for work, empowering them to make well-informed decisions while considering factors such as wages, social security benefits, and taxation?

Addressing such a question could be somewhat challenging at first glance, as it encompasses a range of sub-questions contingent on various factors. Therefore, for the purpose of this assignment, we have chosen to break down the primary research question into three main sub-questions:

1. **Which countries are more favorable based on different factors?**
2. **What are the factors that have a bigger impact on the quality of life?**
3. **How can different countries' indexes relate to each other?**

Splitting up these related questions helps us get a better handle on the main question and find answers more easily. Still, it's important to dive into each indicator separately before looking at them all together to get some valuable insights. For the scope of this assignment, we mainly zoned in on the living cost in European countries, as well as some healthcare indicators.

Firstly, it can be interesting to know at first glance what the average living cost of a citizen in a European country is, since that could be a first general question raised by a graduate about to join the workforce in Europe. The tree map below allows us to better understand the categories that have a bigger or smaller impact in terms of cost of living expenditure.

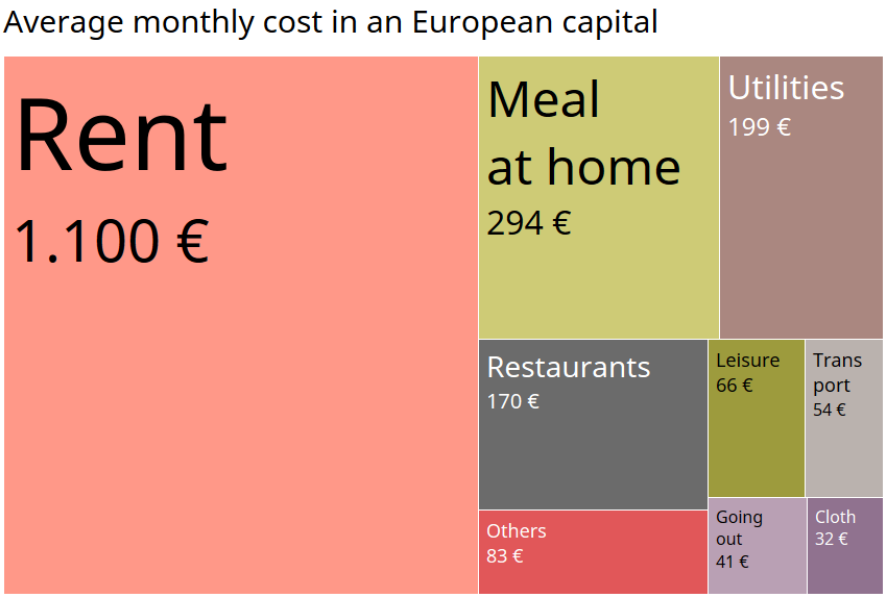
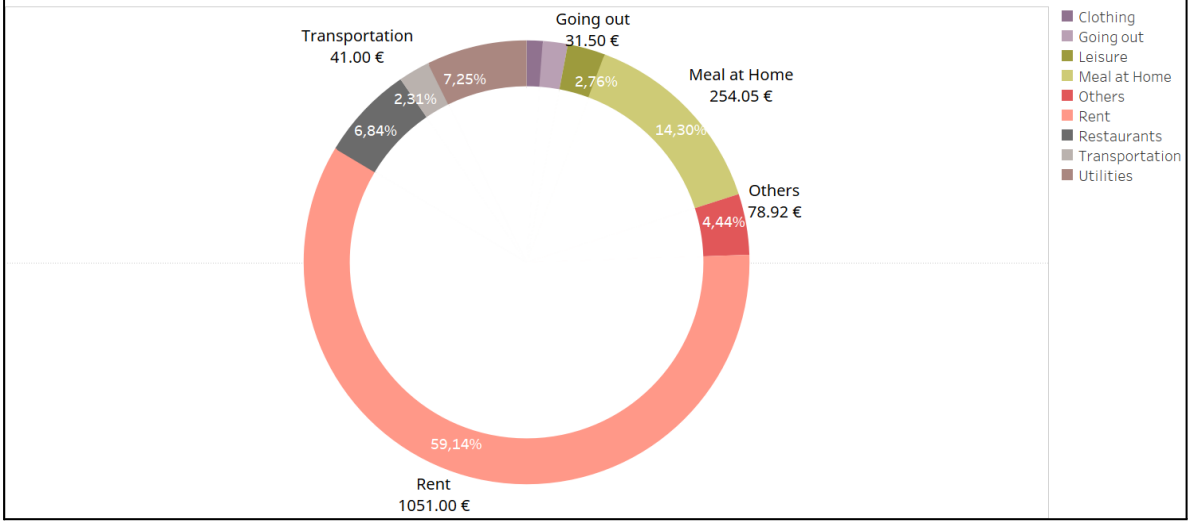


Figure 3 (Daniel Huf)

The visual representation already provides insight into the major contributors when analyzing the cost of living in a European capital, notably highlighting rent, meals, and utilities. It may be beneficial to break down the categories, for example segmenting utilities into water and electricity, and leisure expenses into cinema and sport.

It piques curiosity to understand if this balance remains consistent across all the capitals. The chart below illustrates the distribution of each category with regards to the monthly cost of living in four significant European capitals (the chart does not contain all the capitals of the study for the sake of simplicity).

Monthly Cost Estimator (Madrid 2023)



Monthly Cost Estimator (Dublin 2023)

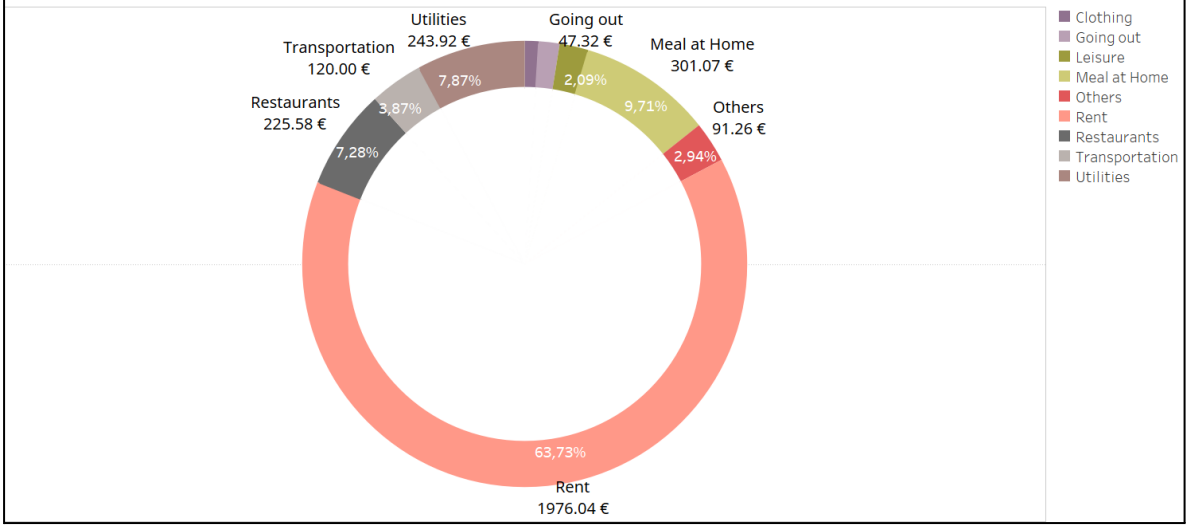


Figure 4 (Daniel Huf)

It becomes apparent that while the relative proportions within each category remain fairly consistent, there exist significant disparities in absolute costs among the capitals, notably in rent and transportation, with Dublin and Zurich exhibiting considerably higher costs in these categories compared to Paris and Madrid. Additionally, Zurich stands out for notably higher prices when it comes to home-cooked meals, a contrast to the other three capitals.

Nonetheless, this chart, in isolation, lacks comprehensiveness since it doesn't encompass all the capitals included in the study, nor does it consider other major European cities beyond the capitals. Despite such raised concerns, the two visualizations presented above do provide a preliminary insight into the expenses associated with residing in Europe. An additional dimension worth exploring in this analysis is the potential impact of purchasing power within each of these cities on the overall cost of living. Such dimension is further explored in the visualization number 12.

Now focusing more on property prices, the graph below shows a time series containing the rent appreciation in European countries, comparing the variations of rent prices inside and outside city centers.

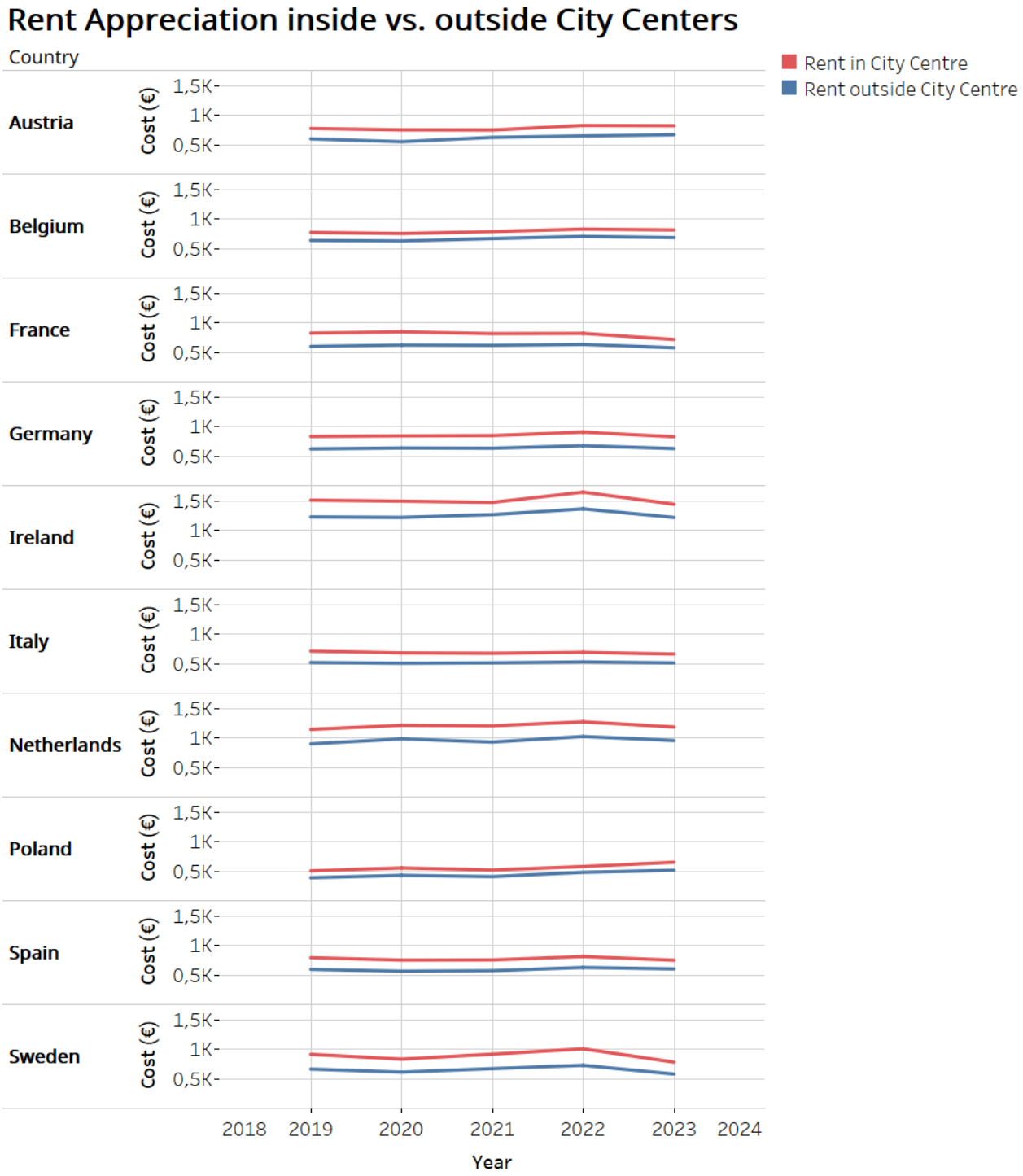


Figure 5 (Daniel Huf)

From such visualization, we can highlight that countries like Netherlands and Sweden present a bigger gap between the rent price inside and outside city centers. In parallel, countries like Sweden and Ireland are characterized by more volatile prices throughout the years. These patterns imply that the decision-making process regarding rental properties should consider not only location but also the smoothness or abruptness of price fluctuations over time.

The property price study can also be tackled by analyzing how different property ratios relate with each other, and how countries differ for each other for the same ratio. The map below shows the price to income ratio that is calculated by following the metrics described below.

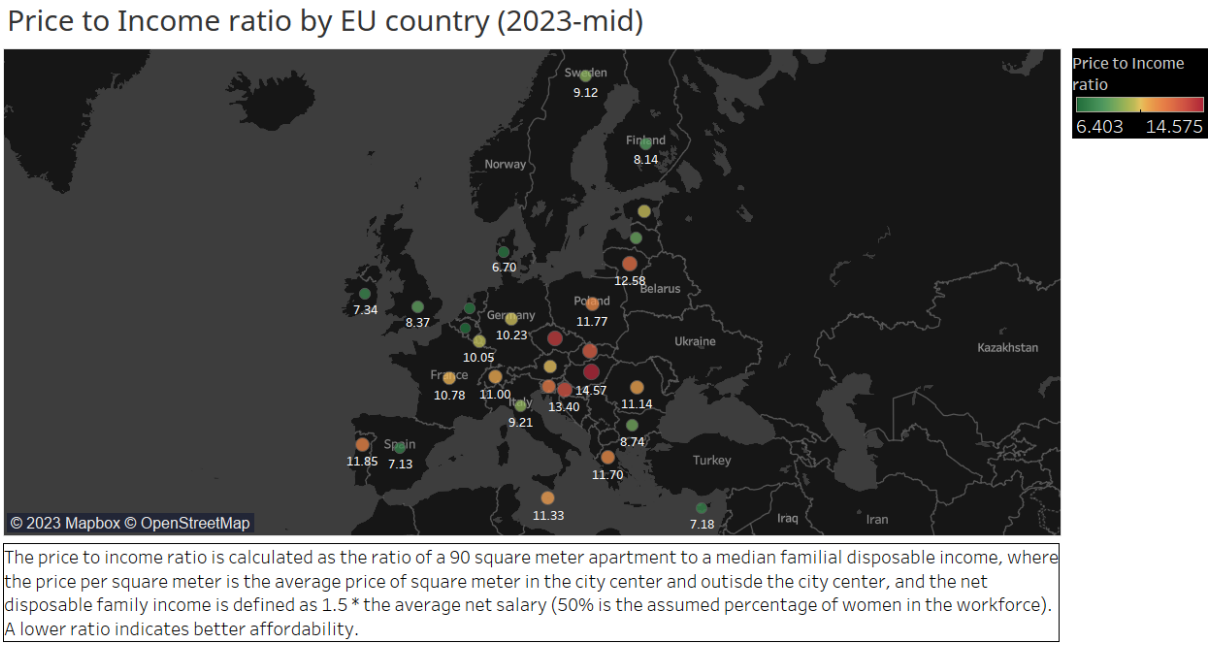


Figure 6 (Daniel Huf)

The visualization suggests some clustered patterns among regions in Europe. Countries located in Central Europe tend to have higher property prices with regard to a person's income, while countries in the West and North propose more affordable prices relative to the workforce's income. Such results can be cross-checked with the price to rent ratio, both inside and outside city centers, which is shown and described in the map below.

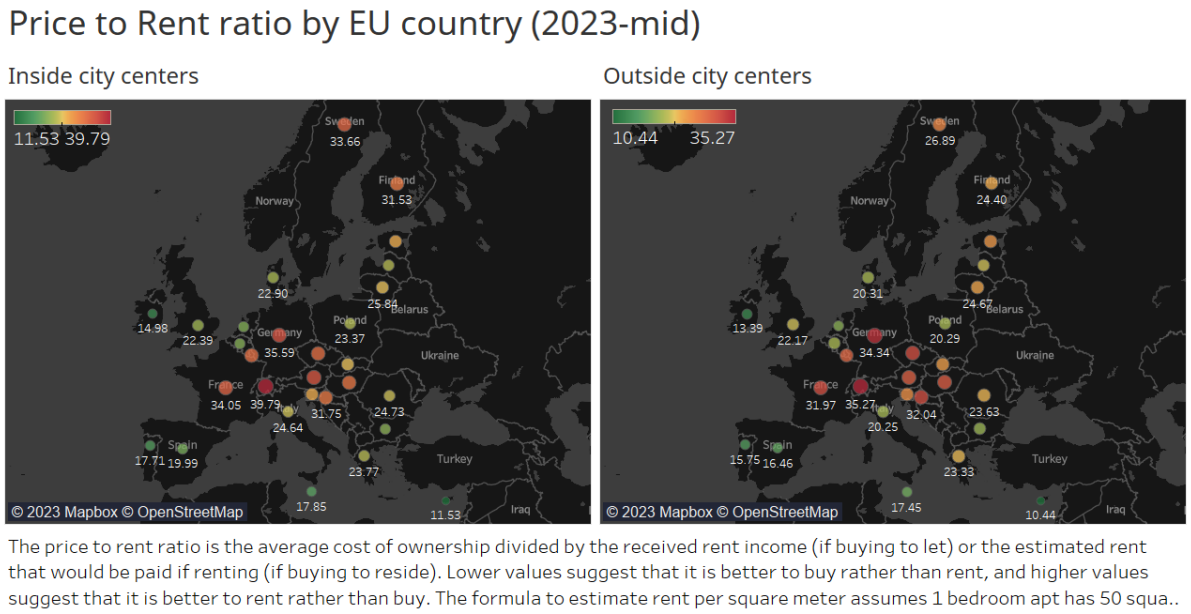


Figure 7 (Daniel Huf)

This visualization suggests that it is better to rent rather than buy properties in Central Europe, aligning with the low affordability in the same region that was presented in the previous map. Conversely, countries in the West suggest that it is better to buy rather than rent, aligning with a higher affordability in the region presented in the previous map.

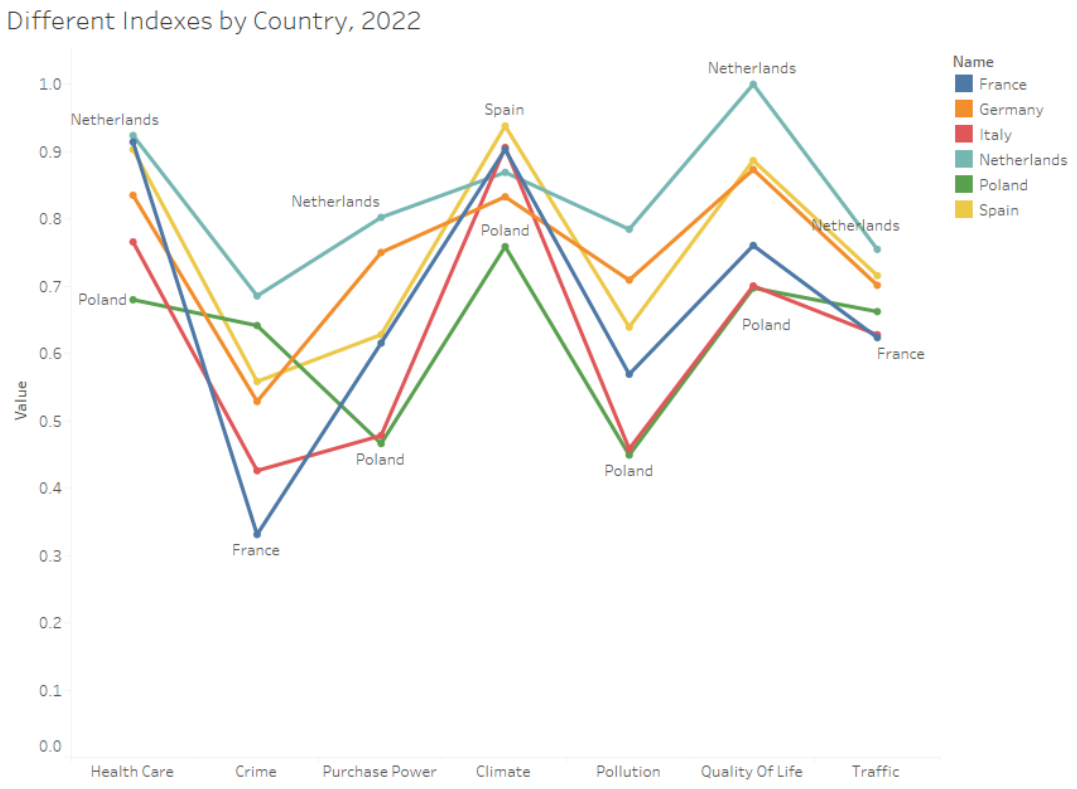


Figure 8 (Sayyor Yusupov)

Next we analyzed the major countries in terms of important indexes. In particular, we used the data for 5 EU countries with the biggest GDP. To obtain this graph, we first had to normalize the indexes to have the same scale. We also modified several indexes to have the highest value as the best (e.g. “Crime Index”). We can see that the Netherlands has the highest value for 6 out of 7 categories, followed closely by Spain and Germany, while the worst performing countries for this data are Poland and Italy. It was surprising that the “Crime Index” has rather low values. Further analysis into crime and other factors might be interesting.

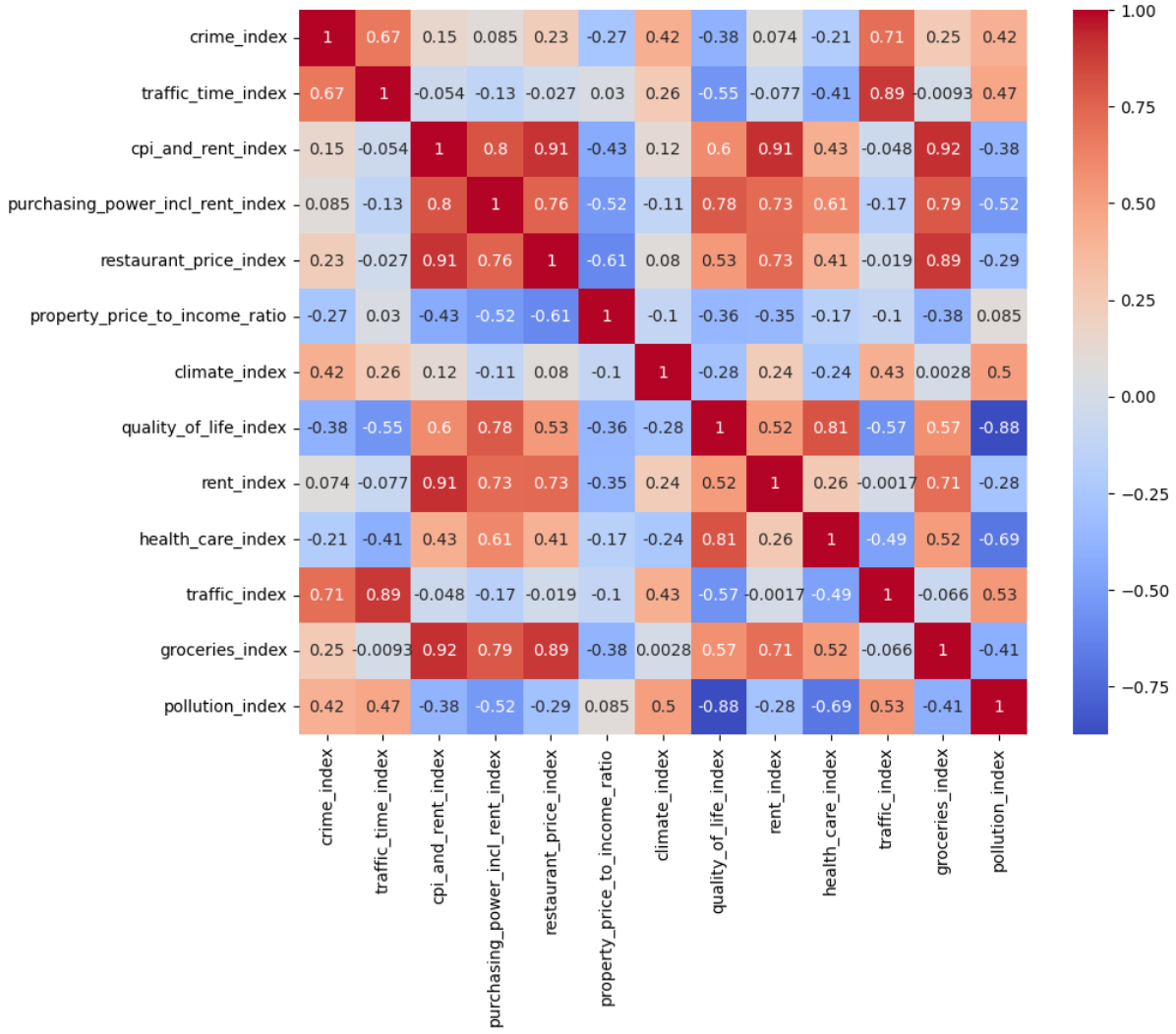


Figure 9 (Mirwise Khan)

Therefore, we compared how correlated indexes, in particular crime, are with each other. Crime is interestingly highly correlated with traffic. One explanation might be that if the government doesn’t handle traffic well, then it’s likely not handling crime well. Also, crime has only a low correlation with quality of life. Surprising correlations for us also included those of quality of life and pollution and health care indexes, health care and pollution. Restaurant price, groceries, purchasing power and cpi are also highly correlated with each other. Thus countries in the EU with higher prices tend to also have higher purchasing power.

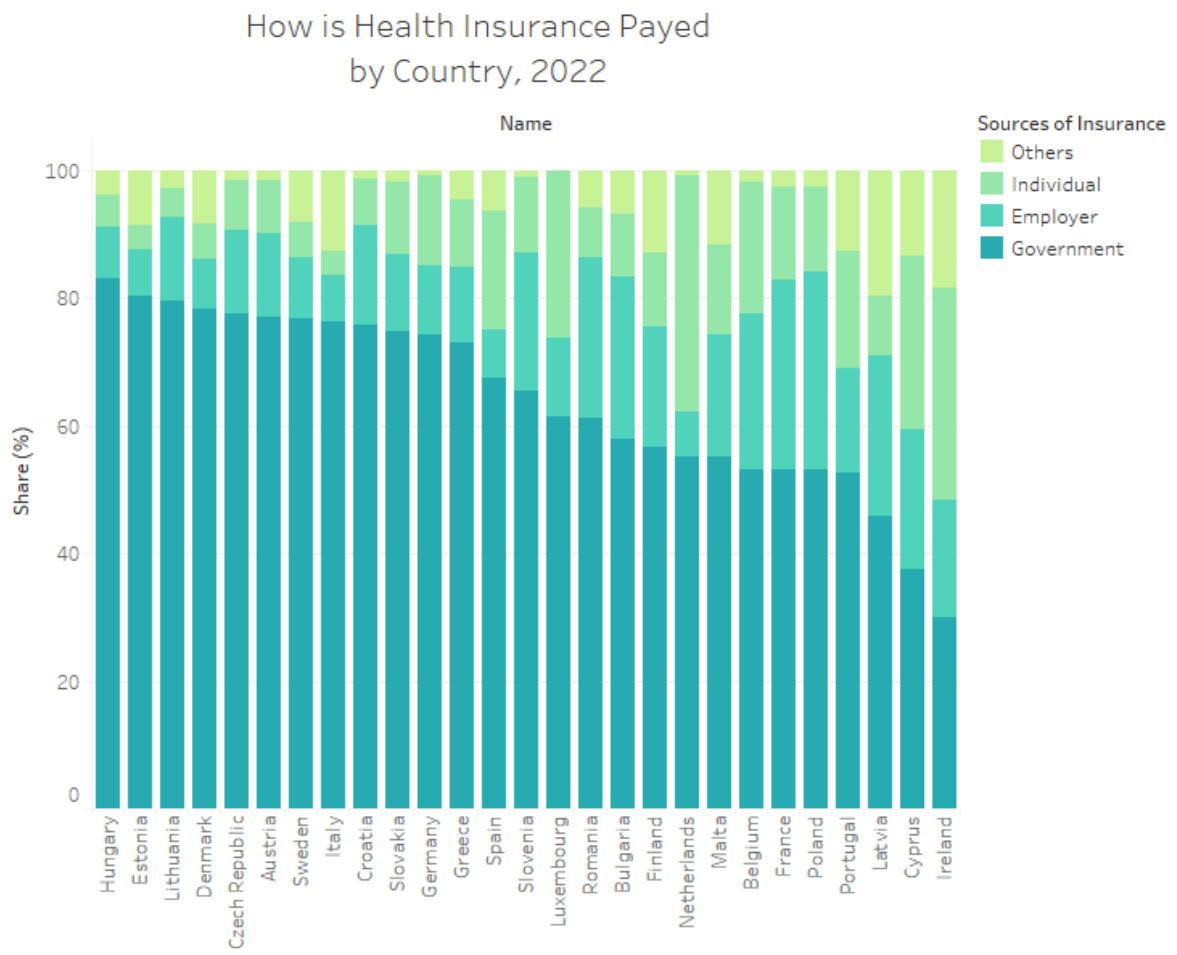


Figure 10 (Sayyor Yusupov)

We also studied the sources of Health Insurance for all EU countries. It is interesting that the governments of 4 Eastern European countries insure the most share of the medical expenses, followed by Denmark and Austria. In general, in the EU the majority of insurance is covered by the government and the employer, with only around 20% of it being covered by the individual.

Choosing a country in the EU to work from a huge list is a challenging task, one of the important factors that can shortlist a set of potential candidates is PropertyPrice-to-Income Ratio as illustrated in the Figure 6. A second important factor can be the commute to the office, and our dataset records information such as **Single One Way Ticket in Public Transport** for each country. Hence, we derived Figure 10.

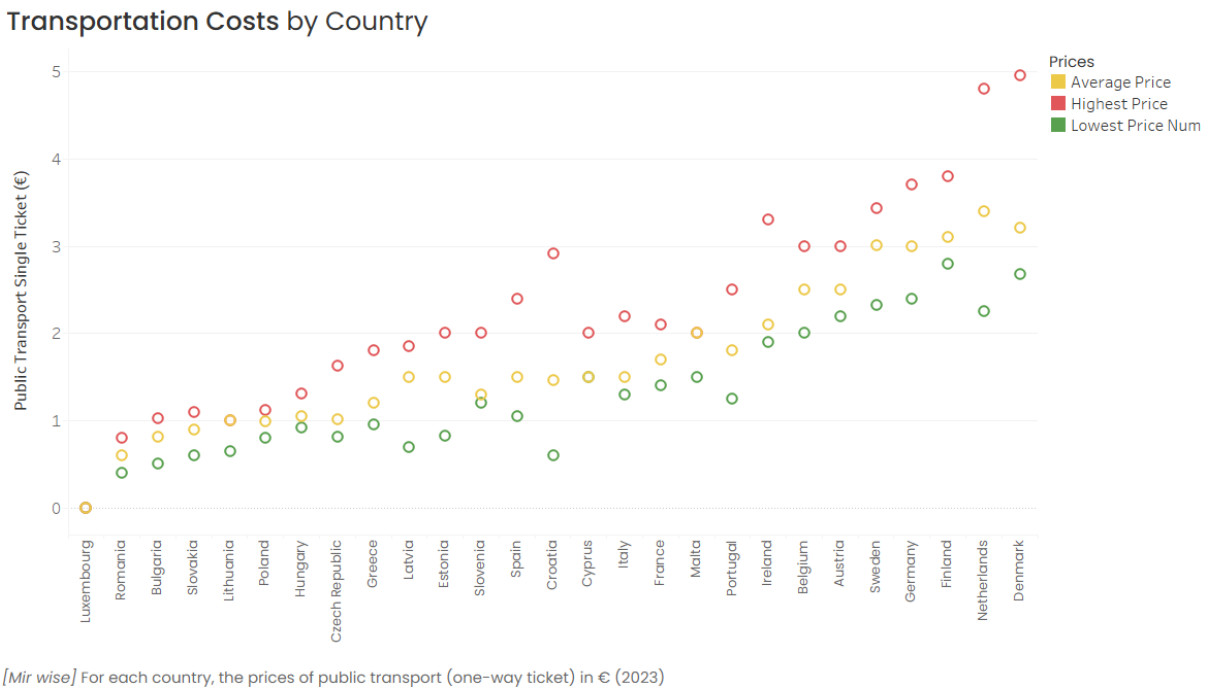


Figure 11 (Mirwise Khan)

Figure 11 serves two purposes, the first being Fact checking - it confirms that the transportation in Luxembourg is 0 i.e. indeed free transportation and the data is indicative of that. Moreover, it doesn't favor Denmark, Netherlands and Finland as they are transportation wise expensive thus favoring Spain and Belgium and potentially Luxembourg as saving transportation/commute cost could compensate for housing affordability factor.

The next interesting variable to assess is the purchasing power of countries and see which ones have remained consistent over the past five years.

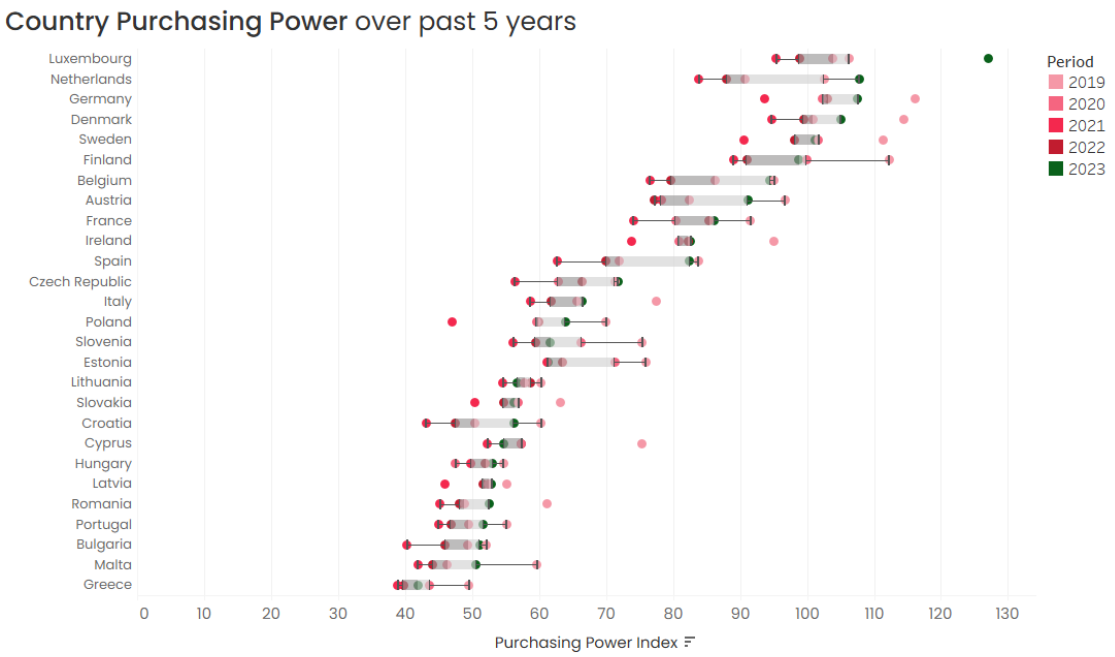


Figure 12 (Mirwise Khan)

Notice this time-series figure was possible to be plotted as a line chart but the whisker's plot highlights the consistency trend more. E.g. we can notice a trend in most cases after 2019 the purchasing power drops and in 2023 (current year) it increases which indicates the economical collapse due to Covid-19 and post-inflation. Furthermore, surprisingly Luxembourg has recovered significantly (Green mark way ahead than Pink), Italy almost recovered while the rest of the countries are still lagging behind their record from 2019.

Summary

In our analysis of the Numbeo dataset, we discovered key insights crucial for selecting an EU country for work. Our findings highlighted important variations in the cost of living among European capitals, emphasizing the significant role of factors such as rent, meals, and utilities. Furthermore, we identified distinct disparities in property prices, rent fluctuations between city centers and outskirts, and regional trends in property affordability ratios. Assessing the quality of life and healthcare data across major EU countries revealed varying performance levels, with some specific countries consistently ranking higher while others displaying areas for potential improvement.

Additionally, our examination of healthcare insurance insights underscored the differences in government support for medical expenses across Eastern and Western European countries. Analyzing the purchasing power trends over time exposed the

impact of the COVID-19 pandemic on the economic stability of these nations, further informing strategic decision-making for individuals considering work opportunities in the European Union.