**Investigating the Global Housing Crisis**

Intro to Data Science – Final Project

Chiara Adjoh-Davoh

ID: 004002088

**Objective**

The purpose of this analysis is to visualize how housing prices, aligned with other economic statistics, have fluctuated and influenced one another, to better understand the ongoing housing crisis in many parts of the world.

**Introduction**

This project uses a dataset originally titled “Global Housing Market Analysis (2015-2024)”. This dataset has 11 columns and 200 rows and provides information regarding housing data for the past decade in various countries. These columns include 1 categorical variable (Country), 1 numerical integer (Year), and 9 numerical floats, including House Price Index, Rent Index, Affordability Ratio, Mortgage Rate, Inflation Rate, GDP Growth, Population Growth, Urbanization Rate and Construction Index.

**Data Wrangling & Cleaning**

The first thing that was evaluated was if the dataset contained any null or missing values. Using the info() function, we can determine how many non-null values there are per column. Since we already know there are 200 rows, 200 non-null values indicates that there are no non-null values in the dataset. It is determined that there are no null or missing values present in the dataset, so we can continue to other aspects of the cleaning process. We also have a built-in function that can drop duplicate rows. When we use this function and then check our dataframe’s shape, we can see that there are still 200 rows after dropping the duplicates, indicating that there were no duplicate rows to drop.

We want to make sure that the values provided to us are sensical and that there was not an error when the data was being compiled. An example of this would be a negative mortgage rate or housing price. When we use the describe() function, we can see the minimum and maximum values of each column, which allows us to see if there are some nonsensical outliers to remove. When examining the minimums of each column, they each made sense and there were no bizarrely large or small values, so we are able to move on to the next phase of our cleaning process.

We now want to rename a few of our variables and drop some others to increase clarity and focus. To start, a variable called “Affordability Ratio” was renamed to “Price to Income Ratio”. The original title of the variable doesn’t make clear exactly what it’s referring to or where the number comes from. When reading the notes for the dataset, I determined that this variable was actually the ratio of the price index to the average income in the area—however average income was not provided in the data set. I choose to give the column this much more precise name instead, which will be useful when it’s time to engineer a new column. Next, I also renamed “Urbanization Rate (%)” to simply “Urbanization (%)”. Rates are defined as changes over time. For example, the Population Growth Rate is a percentage change of the population from year to year. However, when reading the notes for this dataset, it was mentioned that this variable is actually not a year over year change. It is simply a percentage of the amount of people in the given country who live in an urban area at that specific time. This makes it, by definition, not a rate, so I removed the word in order to get rid of confusion. Lastly, I could not find in the notes for the dataset a definition for the “Construction Index” column. There did not seem to be any documentation for what the variable measures or what it comes from. Since it also didn’t seem relevant to the scope of what I was trying to investigate, I removed the column entirely from the data. With the given data cleaned and organized, it was now time to create some of my own columns and features.

**Feature Engineering**

The first thing I created was a categorical variable titled “Region”. This variable simply takes the country and assigns it a region based on its general area. For example, entries with the country, “France”, were given the region “Europe”. While I tried to group by continent, I figured that it would be better to group by more sociological groupings. I also did not want a single country to represent an entire area. So, for example, I created the region “Africa & Middle East” to include South Africa and the UAE. These countries are more socio-politically aligned than, say, the UAE and Japan, despite technically being on the same continent. Another example was grouping Mexico with Brazil, who are technically on different continents but culturally closer together than say, Mexico and Canada.

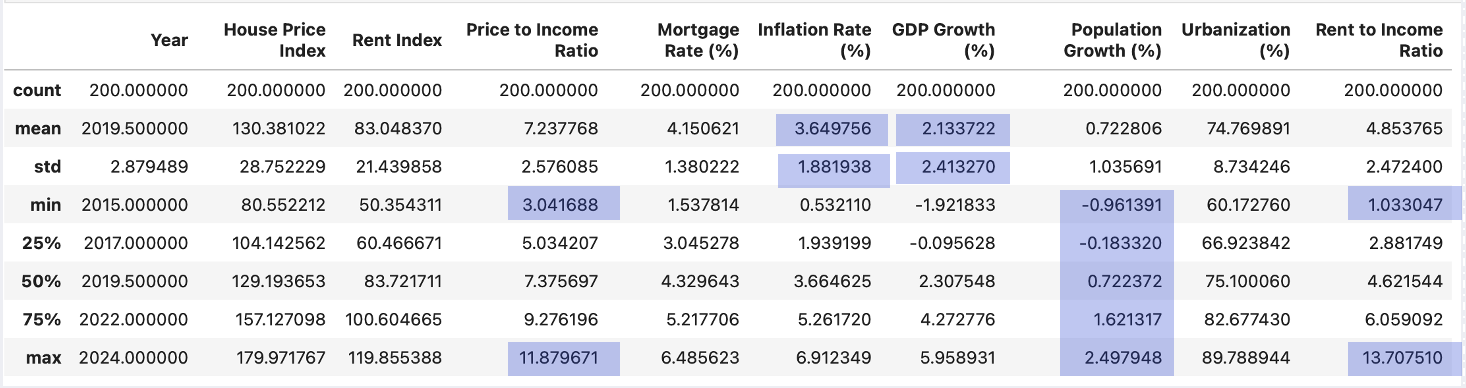
Next, I created a numerical float variable called “Rent to Income Ratio”. This felt like an important thing to measure, but since the average income was not given in the dataset, I had to get creative in order to calculate this. Since Price to Income can be represented as a fraction, Price Index/Average Income, to turn this into Rent Index/Average Income, I multiplied Price to Income Ratio by Rent Index and divided by Price Index. This, hopefully, approximates the Rent to Income Ratio. This ratio represents the affordability of rent for that particular data point, but it can also be a measure of wealth disparity, which will be touched on in a bit.

Lastly, I wanted an easy way to see pre- versus post-pandemic disparities, so I created a categorical variable—represented via an integer—titled “Post Pandemic”. If the year is 2020 or greater, it is given a 1, representing “true”. If the year is before 2020, it is given a 0, representing “false”. These features all proved to be very useful in future analysis and visualizations.

For a later visualization, a variable called “Codes” is also created, with a country code being given to each data point to represent the appropriate country code. This will be used solely for our geographical representations, which we will later review. This is a separate dataframe, so this column will not be present until we begin to look at the geo-graph.

**Exploratory Analysis**

With a cleaned and newly engineered data frame, we can now move to talk about our data from an analytical viewpoint. First, let’s take a look at our summary statistics, obtained using the describe() function. The information we will take a closer look at is highlighted.



The first thing I’d like to touch on is the range difference between the Price to Income (P2I) ratio versus the Rent to Income (R2I) ratio. P2I has a range of 3.04-11.88 while R2I has a range of 1.03-13.71. Rent to Income has a larger range of affordability compared to Price to Income. I’ve interpreted this to be a result of the way that rent has a wider range of affordability compared to house prices. One interpretation of this could be that rent can change much more quickly and without cause compared to house prices. Rent can be increased at the landlord’s discretion after every lease renewal, while house prices are assessed by a third party. House prices tend to change in response to more external factors and can go up and down based on the economy. Another interpretation is that there are a wider range of socioeconomic classes that rent compared to buying a home. Most people who purchase homes fall into a very specific middle-class range, while renting is way more common amongst all parts of the economic spectrum.

Next, the ratio between the standard deviation and mean of the Inflation Rate and GDP Growth is quite high, which I found interesting. A high ratio of standard deviation to mean usually means that there is a wide range of values within the dataset. These two sets are likely facing extreme amounts of fluctuation due to the pandemic and other such socio-political events.

Lastly, I found the differences in the Interquartile Ranges in the Population Growth interesting. There are greater IQRs in the latter half of the population growth than there are in the beginning half which means that more of the data is clustered towards that smaller end. I imagine that a lot of the countries represented in this dataset experienced a population decline or smaller than expected growth due to the pandemic.

After looking at the descriptive statistics, I moved to examine the value counts for countries and regions. When looking at the counts for the countries, there was a perfectly even distribution. Each country had 10 entries in the dataset, one for each year of the decade. However, when we instead look at the value counts for the regions, there was a shocking disparity. Europe alone had 90 entries, almost half of the total 200. The next most region was Asia, with 40 entries, a stunning jump of 50. This means our overall dataset has a bias within it towards European trends. It isn’t necessarily a bad thing, but Europe is not the most populated continent, so it is disproportionate to have it make up such a large part of our global data. When we look at the global trends, we must keep in mind that the data is skewed by the amount of European data, and the rest of our regions may not represent those same trends.

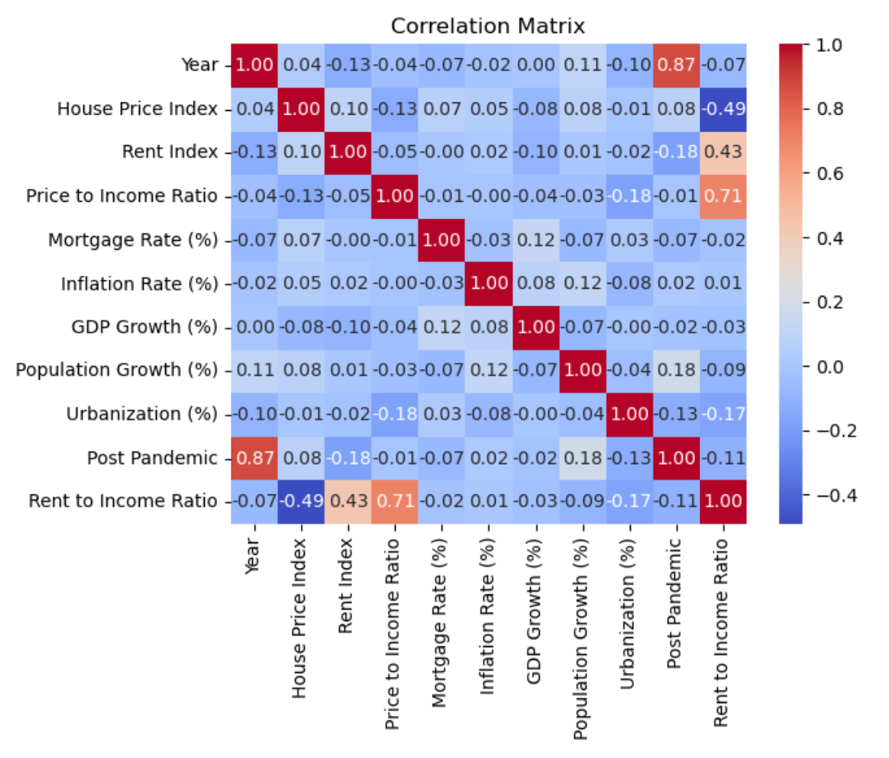
Lastly in this section, I looked at some statistics grouped by region. Firstly, I wanted to see mean P2I and R2I grouped by region. Africa and the Middle East had the highest averages in both categories while South America and Oceania had the lowest. Generally, the P2I and R2I can be interpreted as the affordability of the region. However, this is not the only interpretation of the statistic. A high price or rent to income ratio can also represent wealth disparity. If an area has a small but extremely wealthy class present, these elites can drive up the average rent or home prices without bringing up the average income by all that much. This can result in a situation where most of the homes are not affordable by the average person.

I also took a look at the urbanization percentage for these regions. Asia had the lowest urbanization percentage while Europe had the highest. However, Asia had a higher P2I and R2I than Europe. This led to an interesting possible research topic: does urbanization lead to housing becoming unaffordable? On the surface level, it makes sense. When we think of highly urban areas, such as New York City for example, we associate them with extremely high housing costs because of how densely packed they are. They have a high demand for housing with a smaller supply, so that may lead to housing becoming unaffordable. However, this hypothesis would need further analysis to be confirmed.

**Data Visualization**

It can be hard to understand data just by looking at the raw numbers. An important step in analyzing large datasets is creating visualizations that we can look to.

Let’s start by looking at a heatmap of our correlation index for our numerical values.

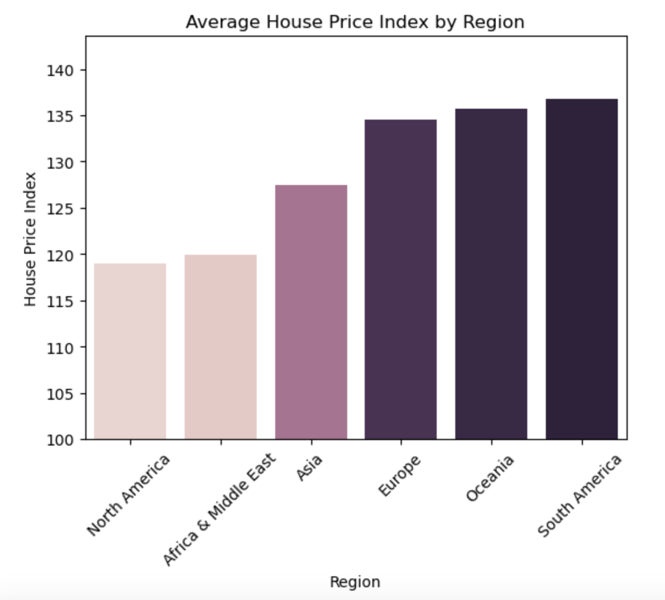
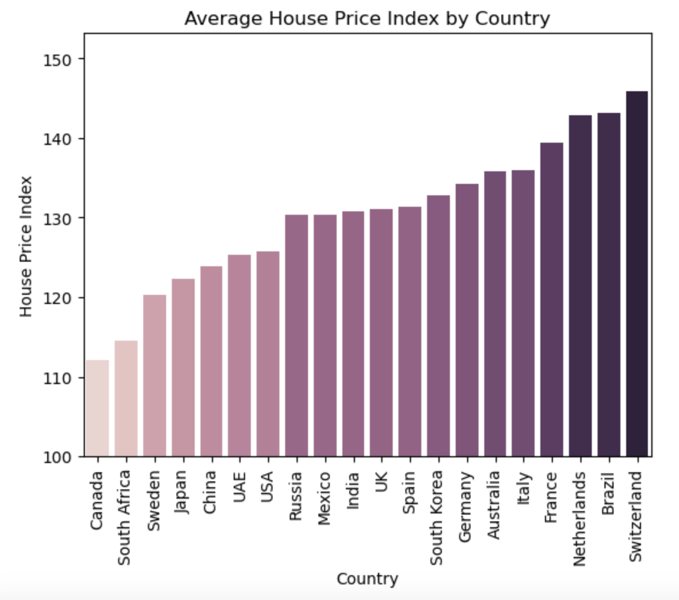


There are a few significant correlations present:

* Rent to Income Ratio vs House Price Index (-0.49)
* Rent to Income Ratio vs Price to Income Ratio (0.71)
* Urbanization vs Price to Income Ratio (-0.18)
* Post Pandemic vs Rent Index (-0.18)
* Post Pandemic vs Population Growth (0.18)
* Urbanization vs Price to Income Ratio (-0.17)
* Rent Index vs Year (-0.13)
* Inflation Rate vs Population Growth (0.13)

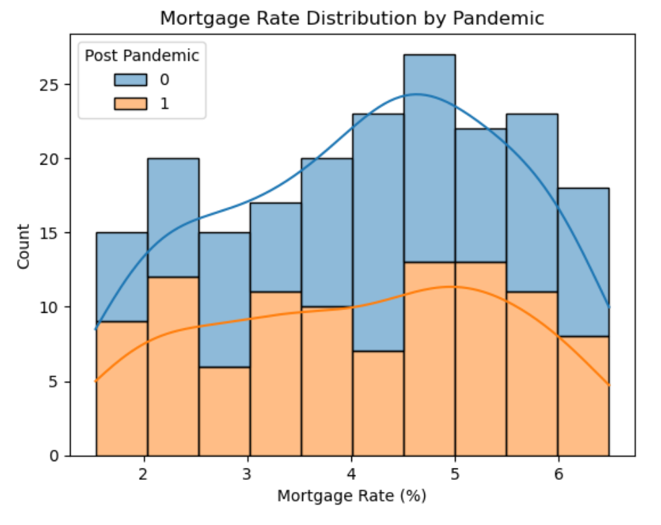
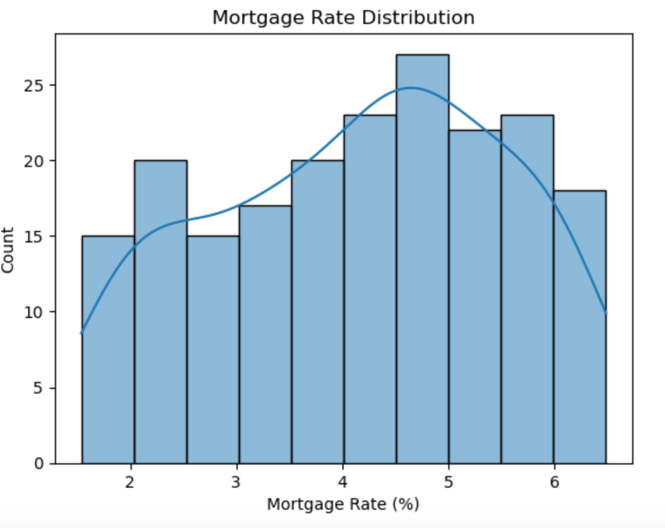
I’ve ignored the correlation between Rent Index and Rent to Income ratio as well as Post-Pandemic and Year because they are calculated using one another. A few of these significant correlations are touched on later but to examine the most interesting briefly: As rent becomes more affordable, houses become less expensive, which makes sense. Cheaper homes can be rented for less. As places become more urbanized, they become less affordable. This tracks, when thinking about supply versus demand. Being post pandemic is inversely correlated with the rent index, so rent is generally cheaper after the pandemic than before.

Moving onto actual visualizations and graphs, I first looked at the average house price index, grouped by both country and region, using a bar plot.



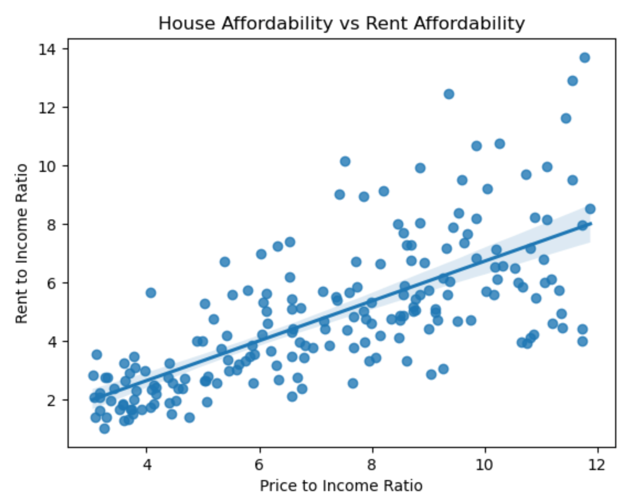
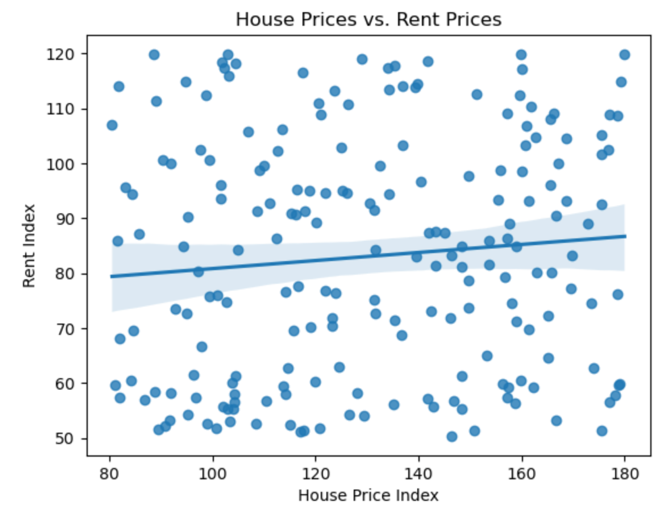
The countries that had the highest house price index were Italy, France, Netherlands, Brazil and Switzerland. 4 out of 5 of those countries are European. However, when we look at the average house price index grouped by region, South America actually has the highest. Europe actually has the third highest. I interpreted this as a result of the European skew of the data. Since there are so many more European countries represented in this dataset, there is a wider variety of values, leading to a more normalized, average value. However, South America and Oceania have two and one countries, respectively, representing them. When there are less data points, it can create an average that is more extreme and less representative of the region as whole. I’m sure if we had more data for South America, it would not be the highest house price index.

Next, I took a look at the spread of mortgage rates, using a histogram. Not only did I look at this overall, but in addition, I also separated the histogram using the Post Pandemic variable to see how it may affect the spread.



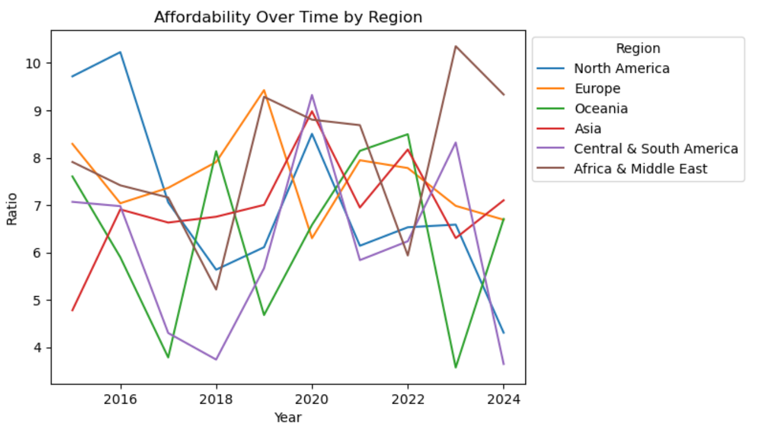
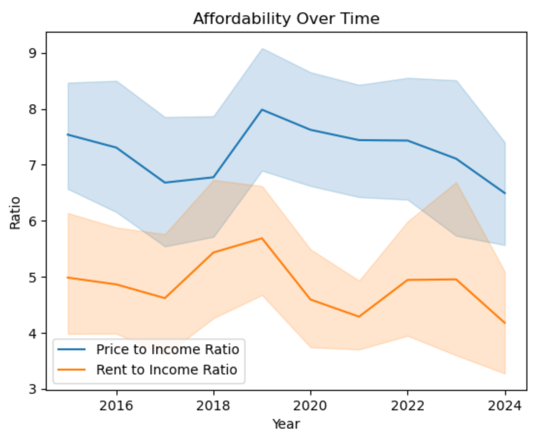
If we take a look at the graph overall, the mortgage rate distribution is slightly skewed left, with a peak in the 4.5-5% range. This is a very middle of the road mortgage rate and does not indicate anything particularly concerning or interesting about the economy. When mortgages are in this range, it usually means that the economy is very stable. However, if we look at this histogram split by the Post Pandemic variable, we see a bit of a different story. While Post Pandemic mortgage rates are still skewed ever so slightly left, there is a far less pronounced difference, and there are a lot more values accounted for in the lower range. For example, in the 1.5-2% range, about 2/3 of the count for this range is from post-pandemic mortgage rates. Many remember that during the pandemic, in the United States specifically, mortgage rates were lowered. However, the 4-4.5% range is about ¾ comprised of pre-pandemic rates. We overall see a much more evenly distributed spread for the rates after the pandemic, representing a sporadic time period where the economy seems to fluctuate.

Next, let’s compare house prices and rent, both in their indexes and their affordability, using a scatterplot with a line of regression.



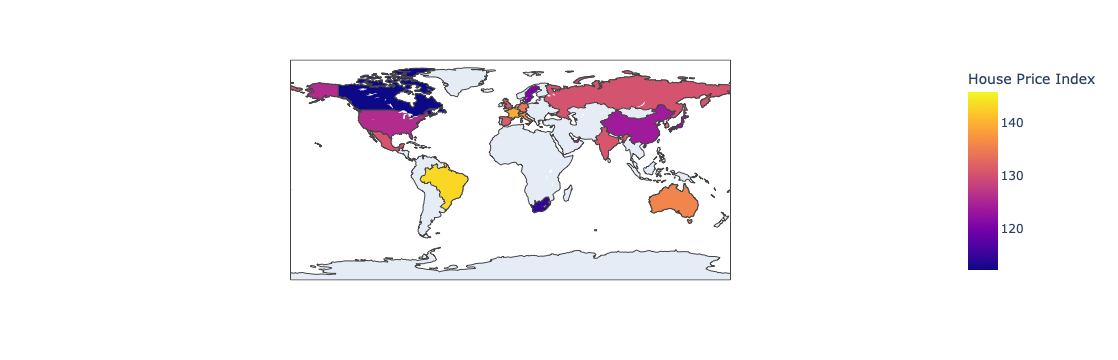
If we first look at just the correlation between the house prices and rent, we do not see a very strong correlation. The data is very scattered with a line of regression with a very flat slope. However, if we take a look at the graph for R2I versus P2I, we see a much more noticeable positive correlation. There are a few ways to interpret this. The first is that while house prices and rent do not affect one another directly, there is likely a third common variable that affects them and causes them to rise in accordance with another. Another interpretation is that since R2I is calculated using R2I, they may be correlated as a result of that.

I next looked at how affordability changed over time, first globally, then by region.



In the first visualization, the affordability of house prices and the affordability of rent were graphed over time. A noticeable trend in both is a peak in affordability in 2019, right before the pandemic, and then a steep and continual drop after the pandemic, leading a record low in affordability in 2024. If we observe affordability, grouped by region however, we can come to a range of different conclusions based on which region we focus on, with none following the same pattern. Oceania, for example, actually peaks in 2022 as opposed to 2019. North America peaked in affordability in 2016, with another relative peak in 2020 before a steep drop for the rest of the decade. Europe most closely follows the pattern shown in the left graph, another representation of how the overall data is skewed by the overabundance of European data.

Lastly, I used the aforementioned country-coded dataframe in conjunction with an aggregated dataframe grouped by country to create a geo map of the average house price index in each country.



This graph is a great way to visualize how the average house price changes region to region, with some places and concentrations being much hotter (higher) than others. It also helps us see trends that are a lot more fluid than just the ones defined by continents or borders. For example, the countries along the same latitude tend to be similar colors. For example, Mexico, India and China. Countries on the same latitude tend to have similar climates, so it may be an interesting question to investigate if climate has an impact on housing costs.

In my project, I also animated the yearly trends for the house price index and the House Price to Income ratio on top of the map. While those animations cannot be included here, when examining them within the code, it is very clear that the geographical trends we discussed before are still relevant. Locations along the same latitude tend to have similar house prices and levels of affordability. Additionally, if we look at how some countries change year to year, we can see that some do rather large fluctuations in short amounts of time. For example, the US changes color quite rapidly in our P2I animation, going through the entire range of colors over the course of our time frame. We also see Russia become more affordable over time, even as the rest of its region becomes less affordable. This can be due to a number of factors, but no doubt the war its embroiled in with Ukraine is a large one.

**Conclusion**

An analysis of the data reveals strong ties to the economic health of a country or region and the affordability of its housing. The pandemic seems to be a driving force behind many of the changes we see in housing costs, as well as the ability for many people to afford the housing that is available. While rent and housing costs may move independently of one another, global economic factors drive the rise and fall of both. It’s always important to include the context of geopolitical events when considering how the economy fluctuates. While the pandemic is the most obvious geopolitical event touched on within this report, there are no doubt a number of events that influenced these changing prices, with some events being more localized than others. Additionally, urbanization is possibly a driving force for the rise in prices and unaffordability, as more people in an area contributes to housing shortages and an inability for people to afford the prices. When we look within this data, we our analysis drives us to encounter even more questions, many of which cannot be answered just by the given dataset alone, especially with such a small group of countries represented. The dataset has its limitations, but it is a useful tool in seeing some localized trends, and it opens the door to deeper dives into the housing crisis.

**References**

Soundankar, A. (2025, March 18). *🏡 Global Housing Market Analysis (2015-2024)*. Kaggle. https://www.kaggle.com/datasets/atharvasoundankar/global-housing-market-analysis-2015-2024

**Acknowledgements**

Thank you to Atharva Soundankar for creating this dataset for my use within this project, as well as Dr. Meenakshi Nerolu for her assistance and guidance within this class.