Project_stat184

Final Project

Introduction and Background

Housing has become almost a luxury rather than a necessity; it is getting harder and harder for people to afford houses. Recently, there has been a drastic increase in house prices due to economic downfall in the United States. It would be beneficial to explore how these prices are affected in completely different countries. The primary goal for this project is to explore the cost- of-living dynamics in each country and how monthly income is related to housing prices in California and how housing prices have been impacted in Japan.

I will be using 3 datasets from Kaggle that contains information about cost-of-living per country,

Japan house prices and California house prices to achieve this goal. Due to lack of datasets and information available online for housing prices in all 50 states, California will be representing housing pricing for the United States; this is purely for comparison purposes.

By exploring the monthly income and housing prices in California, we can provide insights about the country's economy. We can use this to see how "well" a country is doing economically; if more people are able to afford housing it indicates towards a good economical standard. Additionally we want to explore how housing prices have been affected in Japan over the years focusing on 2005 to 2024. Furthermore, It is not just housing prices that we should worry about. The cost of living in each country can affect how high or low a house is priced. It is important to consider the context because the cost of a house can seem reasonable but in reality it is a luxury. A comparative analysis of the cost of living for both countries can provide this context.

Research Questions

1. How do Japan and the United States differ, specifically in terms of cost-of-living changes?
2. How does average monthly income relate to housing prices in California?

3. How do the house prices change over the years in Japan from 2005 to 2024?

Data Summary

Regional Cost of Living (primary dataset) is publicly available on Heidar Sadati's Kaggle page; he collected and updated the data 6 months ago. This data was created for a comparative analysis of income and expense patterns worldwide; mainly used to study affordability and economic well-being across different countries. Each case in the dataset corresponds to a country or region from 2000 to 2023 which includes information about the cost of living index, monthly income, housing percentage, tax, saving percentage, healthcare cost, education cost and transportation cost percentages.

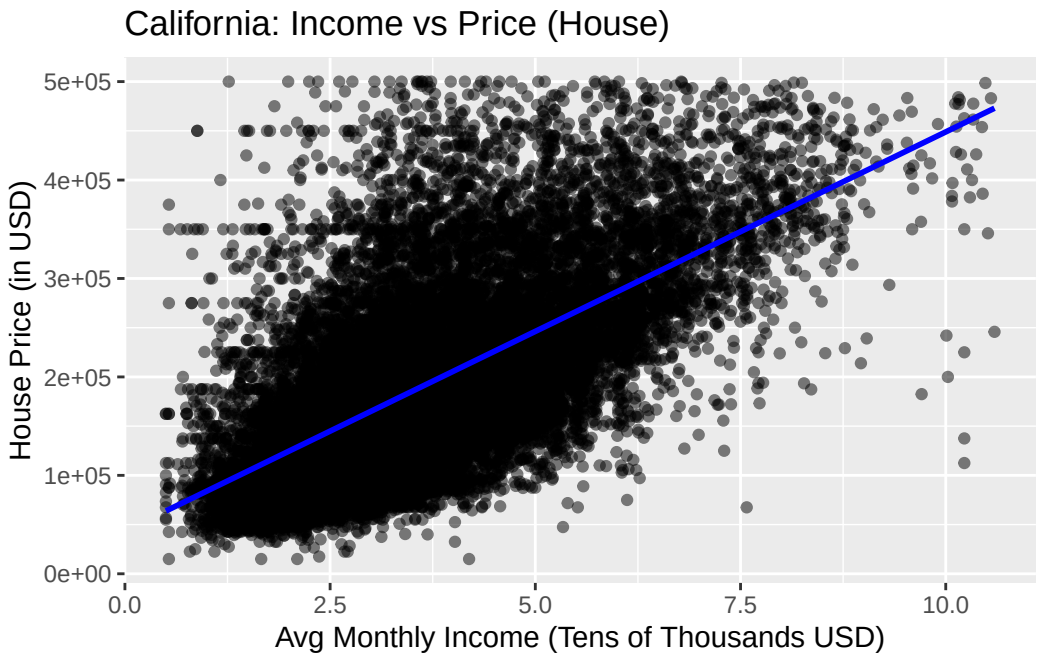
The California Housing Prices dataset is publicly available on Kaggle. The data was derived from 1990 U.S. Census Bureau's survey which was cleaned and uploaded by a user names camnugent. This data was created for evaluating regression algorithms and exploring factors that affect housing prices in California. Each case holds information about the median house value, income, house age, latitude and longitude, count of rooms, bedrooms, and population.

The Japanese Housing Prices dataset is available publicly on Kaggle and is published by Brain McGloughlin. He compiled a dataset using Japan's ministry of land, infrastructure, transport, tourism and real-estate. The dataset was created to develop regression models to forecast Japan housing prices. Each case includes information about type, city, year, average sale price, location and more information about different types of area it is under.

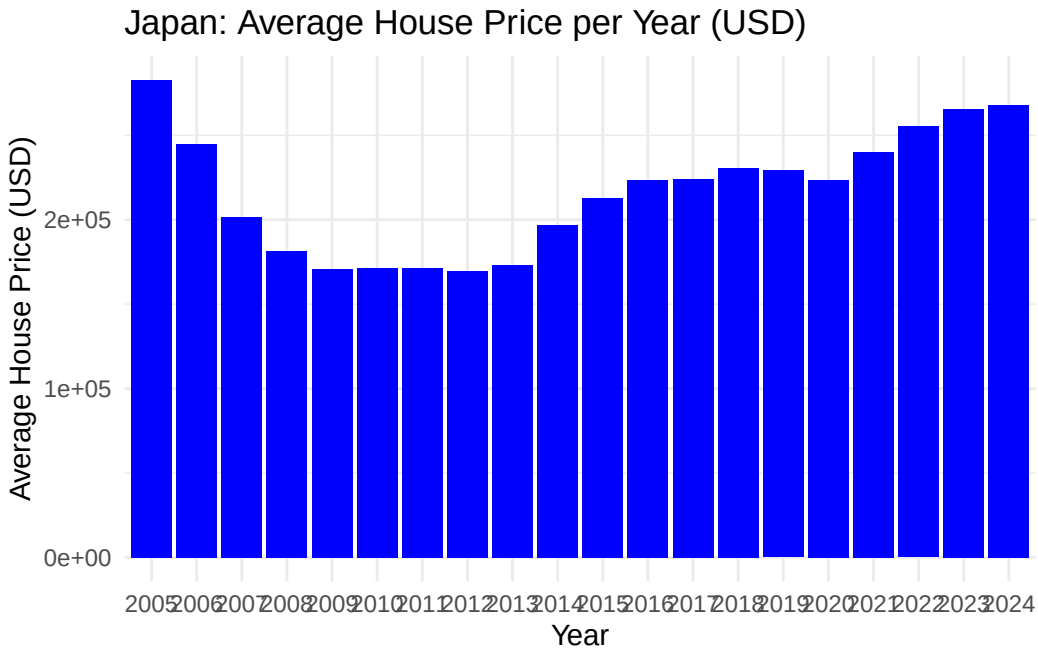
Graphs with Questions

1. How does average income relate to housing in California?

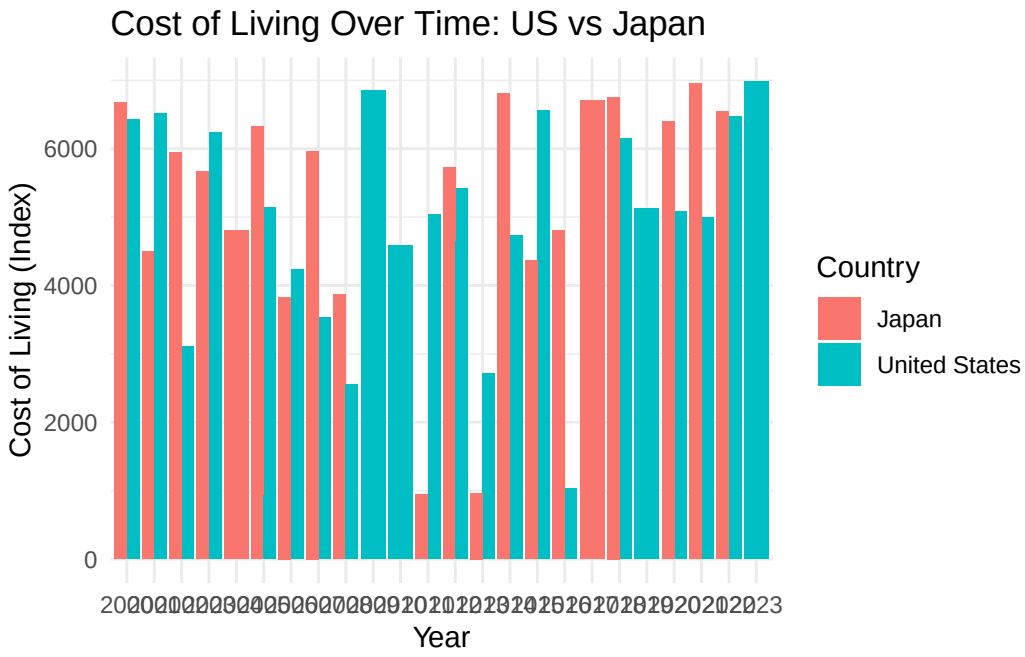
```
`geom_smooth()` using formula = 'y ~ x'
```



2. How does house pricing change over the years in Japan (2005 to 2024)?



3. How did Japan and United States differ in terms of cost of living changes?



EDA

While exploring each dataset, I noticed that the Japan dataset was the largest in terms of data. I had removed irrelevant columns and used the select function to keep only year and total transaction value to explore the price trend over the years. I converted Yen transactions values to USD and created a new column to store USD transaction value using mutate function. I created a bar chart to see how pricing has changed over the years; got rid of any outliers using the filter function.

The California dataset was well collected and did not require much cleaning. I used the select function to create a data frame with only median income and median house columns. To explore how income is related to housing pricing I created a scatter plot with a linear line of best fit; got rid of any outliers for a cleaner scatter plot.

Regional dataset was also well collected and did not require any cleaning. I used the filter function to only have data points for the United States and Japan so I can focus on exploring the cost of living changes in each country.

Conclusion

The upward slope of the blue regression line indicates a moderate positive correlation between average monthly income and house prices in California. There is a dense cluster of data points between 1k to 5k in monthly income and 100k to 300k in house which indicates that most of the population falls within this range.