Name: Housing Market Trends
Author: Vidya Sri Gummadi

Motivation/Rationale for the project: Describe the question you wanted to answer and why is interesting:

Understanding what's happening in the housing market is important, especially when we look at each county in California. This project is all about diving into details like how many houses are being sold for (from low to high prices), who's buying them, and what the economic situation is like in each area. We'll also be checking out stuff like who lives there, their incomes, and how many people are working. Additionally, we'll explore demographic factors such as household composition and enrollment in schools, to understand population trends. By putting all this together, we can help government folks, city planners, and investors figure out where to invest money and how to make housing more affordable. Plus, we'll consider other factors like transportation infrastructure, access to amenities, and environmental considerations to paint a full picture of each county's housing landscape. Basically, we're trying to make it easier for everyone to understand what's going on with houses so we can make our communities better places to live.

Description of data sources: What data did you collect? How did you collect it? How many data samples did you collect? a. Specify exact data sources (e.g., URLs) and your approach to extract the data.

I have collected the data for 3 sectors per county in California which are

1. Housing data - consists of three columns describing the Highest price, Median price, and Lowest Price of all the housing in that county.
2. Economic Data - consists of Gross Domestic Product, Personal Income, Percaptia Income, Labour Force, Employed, Unemployed, and Unemployment Rate.
3. Demographic Data - consists of Population, No. of Students enrolled in the school, Income limits, and Population.

Income Limits consists of Area Median Income(AMI), Acutely Low Income Limit(ALI), Extremely Low Income Limit(ELI), Very Low Income Limit(VLI), Low Income Limit(LI), Moderate Income Limit(MOD)

In total, I have collected 20 attributes from 5 data sources (2 using API, 3 using CSV, and XLSX files)

Here are my data sources and their links:

Economic data(per county):

GDP - datasource1 (API)

Personal income, percapita income - datasource1 (API)
Labor force, employed, Unemployed, Employment rate - datasource2 (XLSX)

Demographic data(per county):

Household composition - datasource3 - (XLSX)
No.of Students enrolled in school - datasource3 - (XLSX)
Income limits - datasource4 - (API)
Population - datasource1 (API)

Housing data (per county):

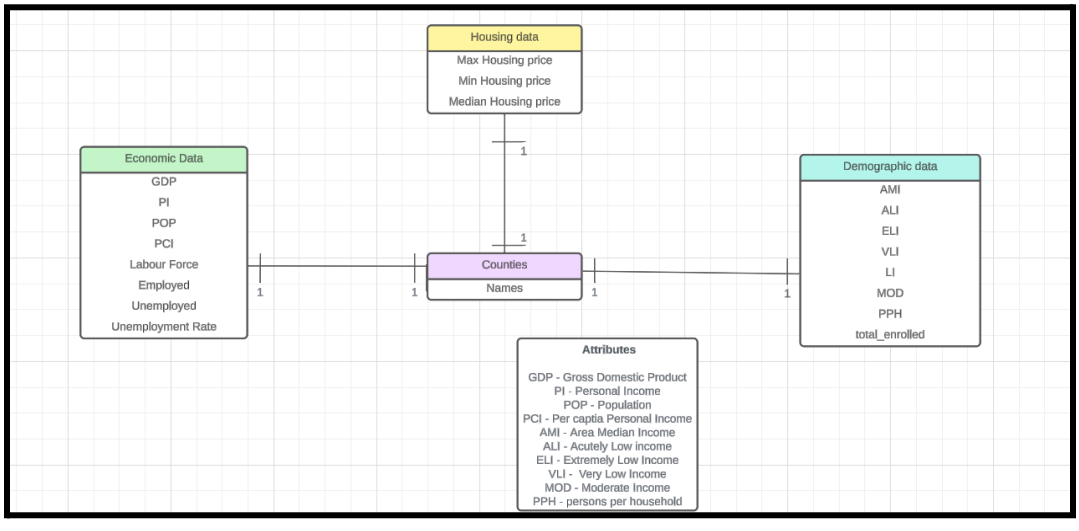
Low price asked: datasource5 - (CSV)
Median price asked: datasource5 - (CSV)
Hig price asked: datasource5 - (CSV)

b. Describe what has been changed from your original plan, and what challenges you encountered or resolved.

Initially, I gathered data on economic factors specific to each county, demographic characteristics covering all states in the USA, and a Housing dataset detailing latitude and longitude coordinates. However, integrating these datasets proved challenging due to the absence of a common merging point. Consequently, I chose to pivot the problem-solving approach, opting to address each county's data individually. This adjustment is anticipated to yield more reliable and insightful results, enhancing the overall understanding of the data landscape.

4. Integrated Data Model: describe the data model and provide an informal entity-relationship diagram. You can reuse the one you provided in submission 2, but update it as necessary if it changes in the final version of the project.

Integrated Data Model: This is my revised version of the ER diagram



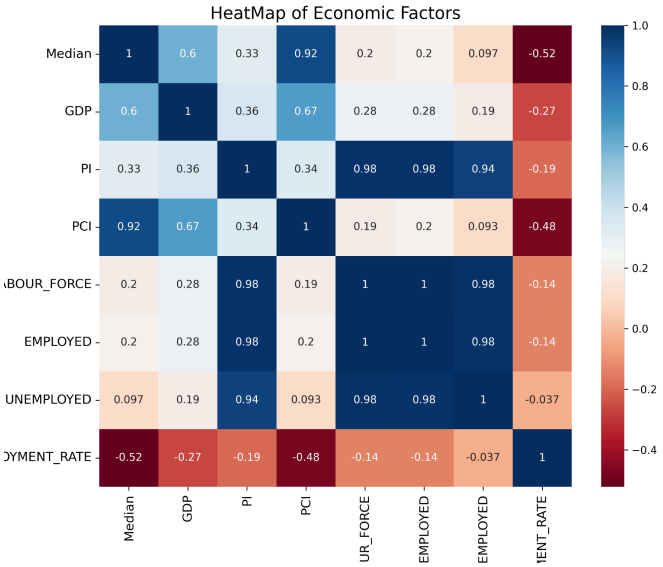
After extracting the data, I cleaned the county-level data from every source and merged them into one data frame for 3 different sectors as shown in ER diagram.

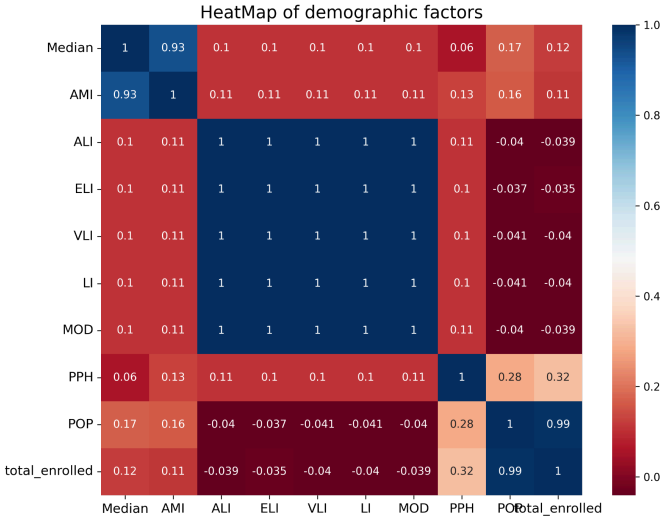
5. Analyses/Visualizations.:a. Describe what analysis techniques you used. b. Describe the figures you made, how you made them, and their elements and meaning.

Initially, I plotted the range of housing prices across all 58 counties to compare how these different price ranges evolve within each county. Specifically, it is a GIF that uses the FuncAnimation function from Matplotlib's animation module to create an animation by updating the plot at each frame.

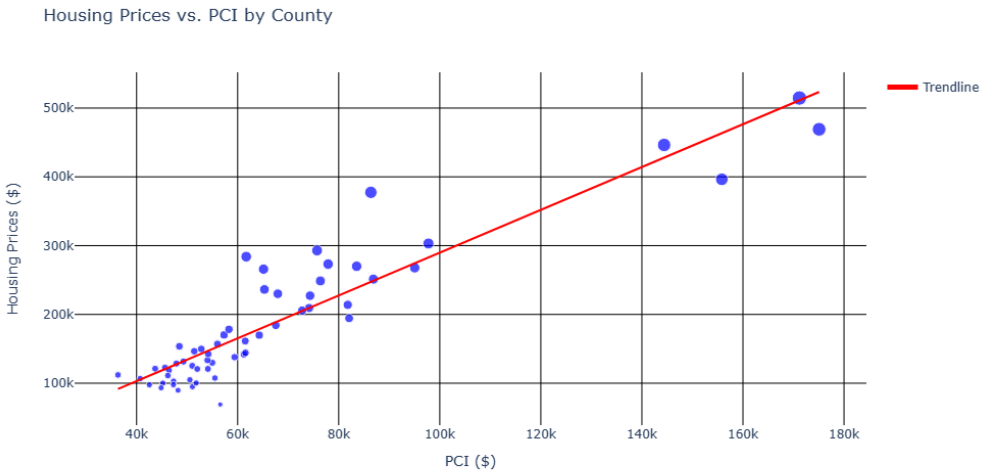
From the above picture, I observed that there are the same trends in price changes for high, median, and low house prices. Therefore, I will go forward with the median house pricing for the next analysis.

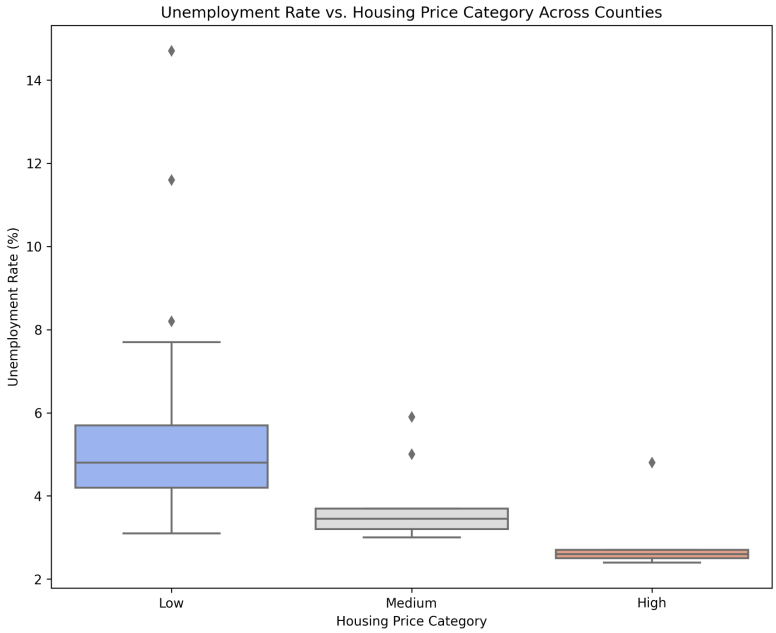
In the next phase, I observed the relationship between the median price and economic factors, demographic respectively, and found the most informative ones using seaborn pair plots which help to identify patterns such as clusters of points, linear relationships, non-linear relationships, or outliers and heatmaps(correlation matrices) providing numerical values indicating the strength and direction of the linear relationship between variables. Analyzing this correlation matrix alongside the pair plots allows for a comprehensive understanding of how the median price relates to various economic and demographic factors. It helps in identifying which factors have the strongest influence on housing prices and guides further investigation into potential causal relationships.

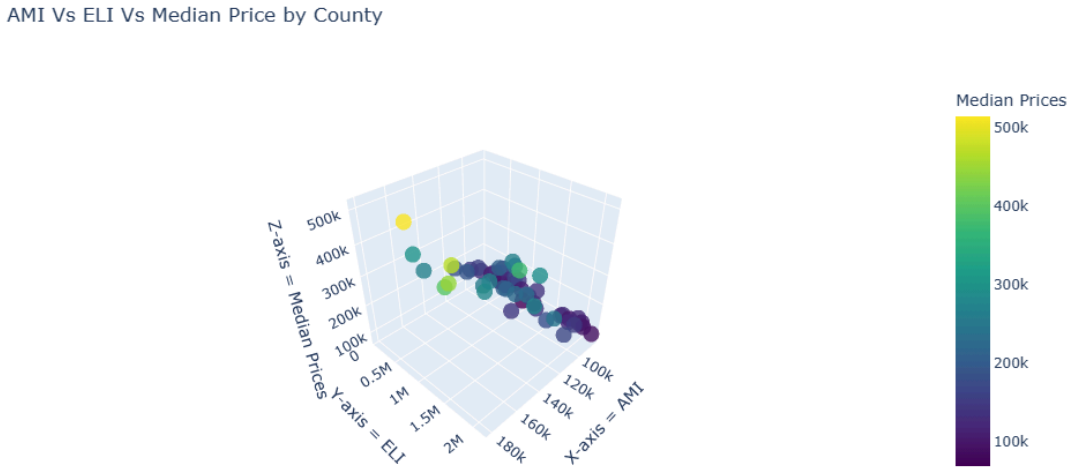

HeatMap of Economic Factors
Median 1 0.6 0.33 0.92 0.2 0.2 0.097 -0.52
GDP 0.6 1 0.36 0.67 0.28 0.28 0.19 -0.27
PI 0.33 0.36 1 0.34 0.98 0.98 0.94 -0.19
PCI 0.92 0.67 0.34 1 0.19 0.2 0.093 -0.48
ABOUR_FORCE 0.2 0.28 0.98 0.19 1 1 0.98 -0.14
EMPLOYED 0.2 0.28 0.98 0.2 1 1 0.98 -0.14
UNEMPLOYED 0.097 0.19 0.94 0.093 0.98 0.98 1 -0.037
OYMENT_RATE -0.52 -0.27 -0.19 -0.48 -0.14 -0.14 -0.037 1
Median GDP PI PCI UR_FORCE EMPLOYED EMPLOYED IENT_RATE
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4


HeatMap of demographic factors
Median 1 0.93 0.1 0.1 0.1 0.1 0.1 0.06 0.17 0.12
AMI 0.93 1 0.11 0.11 0.11 0.11 0.11 0.13 0.16 0.11
ALI 0.1 0.11 1 1 1 1 1 0.11 -0.04 -0.039
ELI 0.1 0.11 1 1 1 1 1 0.1 -0.037 -0.035
VLI 0.1 0.11 1 1 1 1 1 0.1 -0.041 -0.04
LI 0.1 0.11 1 1 1 1 1 0.1 -0.041 -0.04
MOD 0.1 0.11 1 1 1 1 1 0.11 -0.04 -0.039
PPH 0.06 0.13 0.11 0.1 0.1 0.1 0.11 1 0.28 0.32
POP 0.17 0.16 -0.04 -0.037 -0.041 -0.041 -0.04 0.28 1 0.99
total_enrolled 0.12 0.11 -0.039 -0.035 -0.04 -0.04 -0.039 0.32 0.99 1
Median AMI ALI ELI VLI LI MOD PPH POP total_enrolled
1.0 0.8 0.6 0.4 0.2 0.0

After this, I fitted the simple linear regression using stats models using all the economic factors and demographic factors respectively, and visualized some significant variables with a scatter plot with a thread line using Plotly Express (plotly.express) library (also specifies the content of every datapoint when hover on to it), and the trendline is added using Plotly Graph Objects (plotly.graph_objects) library, simple box plots, and 3D scatter plots using Plotly Graph Objects with interactive features include hover information, which displays detailed data when hovering over data points, and a color scale legend indicating median prices. Additionally, the plot is fully rotatable and zoomable .







At last, I again fitted the regression model with all the potentially significant variables.

6. Conclusions: Describe your findings and their impact.

From **Figure 1(Profile Map of Housing Prices across all Counties)**: I observed that there are the same trends in price changes for high, median, and low house prices. Therefore, I had selected the median housing price for the rest of my analysis which is more reliable.

From the pair plots and correlation matrix of economic factors, I observed that percaptia income(PCI) has a positive correlation with the housing price and also the unemployment_rate has some significant effect of negative correlation with the housing price which means as the PCI increases the housing price also increases and combining if the unemployment is also less for that county, it will have high housing prices.

I have visualized these findings through **Figure 5 boxplots for unemployment_rate showing a negative correlation** and through **Figure 4 scatterplot with thread-line for PCI(Per Captia Income) showing a positive correlation**. But surprisingly, none of the other economic factors plays a role in predicting housing values

From the regression analysis, I found how much amount of these variables are significant from the rest all the economic factors.

From the pair plots and correlation matrix of demographic factors, I observed that Area Median income(AMI) has a positive correlation with the housing price but after using regression analysis, I found that Extremely Low Income Level also has some significant importance towards housing price. I have visualized these findings through **Figure 6 3D scatter plot.** None of the other demographic factors plays a role in predicting housing values which is surprising.

From the regression analysis, I found how much amount of these variables are significant from the rest all the demogrpahic factors.

7. Future Work: Given more time, what direction would you take to improve your project?

Given more time, I will try to develop predictive models to forecast future housing prices based on a combination of housing, economic, and demographic variables. Machine learning techniques such as regression analysis, random forests, or gradient boosting can be applied for predictive modeling.And also considering external factors such as crime rates, race/ethincity, interest rates, government policies, and local regulations affecting the housing market. Assessing how changes in these factors impact housing supply, demand, and prices.