Introduce the problem
The question that I want to answer with this data is how much the price of houses sold has increased over the years based on the particular state being looked at.

Introduce the data
I found the following 2 of my datasets from Kaggle because I couldnt find any datasets from other data websites. In order to combine both of these data sets I wil need to base it off of the address and price sold because I can try to add the second dataset into the first dataset (append).

- https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset
    - This dataset contains information about real estate such as status, price, address, house size, and sold date
- https://www.kaggle.com/datasets/crawlfeeds/redfin-usa-real-estate-data
    - This dataset has price, address, and property details like bed, bath, and square feet.

I also found this dataset from the US Census Bureau website:

- https://www.census.gov/topics/housing/data/tables.2021.List_395043788.html
- Annual_AvgPrice.xlsx
    - This dataset contains Average Sales Price of New Manufactured Homes Sold based on each state. I will not be using this dataset until the very end after filtering and processing the other 2 sets.

Pre-processing the data
The steps that I followed to pre-process my data:

- Getting rid of irrelevant data and only keeping columns such as price, address, state/region, and sold_date/scraped_at in each dataset.
- Dropping null values to get a better understanding of valid data. The first dataset had 466763 values that were null for sold_date and 71 values that were null for price. The second dataset had only 1 null value which was for the price.
- I standardized the data to make sure both datasets had the same column names. This made it more cohesive. This will also make it more convenient to combine both datasets later on.
- I converted the date sold object to a datetime type so that it is easier to extract the year from the date for both datasets. Then I actually extracted the year from that column and just reassigned it back to date sold.
- I also used scaling to reset the index numbers after removing irrelevant data.
- Since the first dataset had the states as the full name instead of the 2 letter abbreviation, I changed that to match with the second datatset.
- I had to remove all null rows that resulted from changing the states to 2 letter abbreviations because some states were not located in the US.
- I got rid of the more irrelevant data such as the address because I found no use for it after funneling my data.
- I made sure to make my jupyter notebook easy to understand and added comments.

Data Understanding/Visualization

- In order to understand the data I tried to visualize it in many ways. There was a lot of trial and error.
- Firstly, I wanted to do a geographic map where you can see the different states and when you hover over the state you'll be able to see the average price of the houses that were sold for a particular year. Unfortunately, this did not work as planned because I wanted to see the data for more than just one year at a time.
- Then after having my work reviewed by my peer, I saw that there was a better way to visualize the data and that was in the form of a line plot with column encoding color. This was a good way to show all three components of Year, State, and Price but it got really messy and was not as organized as I would have liked it to be. I was hard to separately see the different years.
- Then I started looking at heatmaps which seemed to be very organized and clear to see all the information. This was quite challenging to create and code because it was a lot of different components and I was having a lot of errors and it took me a long time to figure out the syntax and incorporate my data into it.
- The heatmap kept giving me errors so I had to settle for a pair plot. This did not show much of what I was looking for but it was the only one that worked with my data.
    - After visualizing the data with this pair plot, I saw that there was an outlier in the range 2001-2003. I did some trial and error to remove this outlier for a better visualization. After removing it, it was easier to see the rest of the data.

Storytelling

- I learned that my assumption about house prices have increased over the years. Although there were some up and down in the visualization (the most significant down being in 2010 which is when the housing market crashed). I believe the data to be pretty accurate in regards to the prices in different states as seen in dataframe 2 as well as seeing a corelation in years and prices in dataframe 1. Overall, I was able to answer my original question.

Impact Section

Discuss the possible impact of your project. For example, how could your visualizations cause possible harm? What data or perspectives might be missing from this work?

- The answer to the question can be useful to predict the market trends and to see if there is any correlations between intervals of time and house prices. Information such as inflation and market crashes will also need to be looked at separately in order to fully get the answer but that information will be missing from the initial data analysis that I will be doing based on the dataset.

References

- https://www.datasciencemadesimple.com/get-year-from-date-pandas-python-2/
- https://stackoverflow.com/questions/66572349/python-sub-state-names-for-abbrev-via-python-dict-with-re-sub

- https://www.kaggle.com/code/robikscube/baby-name-popularity-eda#Name-Popularity-by-State
- https://datavizpyr.com/heatmap-with-matplotlib-in-python/
- https://matplotlib.org/stable/gallery/showcase/bachelors_degrees_by_gender.html#sphx-glr-gallery-showcase-bachelors-degrees-by-gender-py # didnt use but might in the future
- https://cmdlinetips.com/2021/04/convert-two-column-values-from-pandas-dataframe-to-a-dictionary/

Code

- https://github.com/ChernyDevireddy1/ITCS3162/blob/main/Project%201%20Notebook.ipynb