CSC 251 PROJECT 3

HOUSING MARKET ANALYSIS

THE ASSIGNMENT

The US has observed a huge increase in prices over the last few years. Home prices, rent, groceries, construction - just about everything now costs much more than it used to. Blame it on inflation, the pandemic, mortgage rates, democrats, or republicans (or all of these), the cost of living has soared. In this project, you will look at the increase in housing costs over the last 25 years and how that increase compares with the increase in rental costs, college tuition, and the Consumer Price Index. You will gain experience in investigating and comparing datasets, manipulating and merging DataFrames, visualizing data and computing metrics. Your notebook should “tell a story” about the data and your observations in a well-organized, easy-to-read format.

DATASETS PROVIDED

The following datasets are provided for you as .csv files:

1) **Metro_zhvi_homes_smoothed_2025.csv** – The Zillow Home Value Index (ZHVI) refers to average home prices for mid-range homes in the 35th to 65th percentile range for hundreds of US metropolitan areas between 2000 and early 2025 gathered from https://www.zillow.com/research/data/ . This data set is smoothed to average out seasonal fluctuations in home prices. You will be most interested in three sets of data: prices for Raleigh, NC, prices for Greensboro, NC, and prices for one more market outside of North Carolina of your choice. Pick a market that may be significantly different from the two NC markets to make it more interesting!

2) **college_tuition_data.tsv** – This file is tab-delimited and contains average undergraduate education costs for all public and private US institutions from 1963 through the present. I’ve already done some preparation on this file so that it imports cleanly into your notebook. Since some of the column headings have disappeared, you’ll want to visit https://nces.ed.gov/programs/digest/d23/tables/dt23_330.10.asp?current=yes so view the table and be sure to select the column for “Tuition and required fees: All institutions” in **current** dollars.

3) **Metro_zori_monthly_rentals_smoothed_2025.csv** – The Zillow Observed Rent Index (ZORI) refers to the market rental rate for properties within the 35th to 65th percentile range for all homes and apartments for hundreds of US metropolitan areas between 2015 up through early 2025 gathered from https://www.zillow.com/research/data/ . This dataset includes all types of rentals including apartments, condos, and single-family homes. You will focus on the same 3 markets you considered for home values.

4) **cpi.csv** – The Consumer Price Index (CPI) is a calculated measure of inflation released monthly from 1913 to 2024 by the US Dept. of Labor Bureau of Labor Statistics. It is based upon the prices of goods such as food, clothing, shelter, fuel, taxes, transportation costs, and service fees for urban consumers. This data set was gathered from https://fred.stlouisfed.org/tags/series?t=annual%3Bcpi .

For each file provided:

1) Import as a .csv or .tsv file and save the data in a DataFrame.
2) Complete all appropriate wrangling of the DataFrame within the notebook. This may include:
    a. Filtering and subsetting the DataFrameto select the columns and rows of interest
    b. Removing observations with missing values
    c. Transposing or pivoting the data
    d. Renaming columns or indexes
    e. Converting the format of dates
    f. Grouping rows of data to perform aggregate calculations
3) Merge datasets into the same DataFrame prior to performing statistical analysis or creating visualizations.

DATA ANALYSIS

Your analysis should include (not necessarily in this order):

1) A univariate description of each row/column of data in written and graphical form. This includes time-series data graphed as a function of time.

2) Bivariate analysis such as scatter or line plots, correlation, statistical correlations, and numerical comparisons. Since inflation and other factors drive prices up, we don't need to run regression models and generate equations to predict the increase in one value based upon another. We can, however, compare the rates of increase between two datasets.

    Tip: When making comparisons, consider the Pandas pct_change function. This handy function calculates the fractional change between subsequent row values in a DataFrame, leaving the first row with a NaN. Don't forget to multiply by 100 to get the actual percent change. For this analysis, it may be appropriate to find an average monthly percent change then multiply by 12 to get an average annual change. (While this may be slightly different than reality since the actual change is compounded over time, it's a reasonable measure of comparison and compares nicely with published percent increases.) More information can be found online including the following site: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.pct_change.html

3) An attempt to answer the following questions (and similar questions like these) as you explore the data:
    a. How do the 3 housing markets compare (graphically and numerically)? What patterns do they have in common? Do you observe many outliers? Any features worth commenting on? Any speculations on the patterns or features? What was the reasoning behind your pick for the 3rd market? How is it different from the other 2 markets?
    b. How does the increase in tuition and fees at US colleges compare with the increase in the 3 housing markets? Any patterns or strong correlations?
    c. How do the 3 rental markets compare with each other and with the home prices? Any patterns that you observe?
    d. How does the increase in Consumer Price Index compare with the housing markets? What patterns do they have in common? Any strong correlations?

4) A brief summary of your observations. Consider a bulleted list rather than paragraph after paragraph rehashing everything all over again.

5) A reflection of what you learned about working with data, what was challenging, and what didn't work. Once again, not an essay but a few short paragraphs or a bulleted list.

PROJECT REQUIREMENTS

1) Reminder: This is an individual project so make sure to complete your own work! Be sure that your project and your neighbor's doesn't begin to look the same!
2) Submit a Jupyter notebook that includes all dataset descriptions, data manipulations, visualizations, correlations, conclusions, and reflections.
3) Your Jupyter notebook should include a markdown cell that includes the project information (name, date, project name and description, etc.) at the top of your file and markdown cells throughout your notebook with subtitles, conclusions, and comments arising from your analysis. Your summary and reflection should each be labeled in separate markdown cells.
4) The datasets provided should be imported into your notebook in their given .csv or .tsv format and manipulated within the notebook.
5) Only the following Python packages may be used: pandas, matplotlib, seaborn, statsmodels, and datatime. Also, if you wish to hide any warnings, you may use the script:

```
import warnings
warnings.filterwarnings('ignore')
```

6) All visualizations should include a title, axis labels, legend, and appropriate scaling.
7) Your notebook should be cleaned up and extraneous code should be removed. Reduce clutter by combining commands within the same cell or not showing the manipulated DataFrame after each command in your final notebook. You are encouraged to include multiple plots on the same axes and multiple axes in the same figure provided they are effective representations of the data. Resizing figures may also assist in preventing your notebook from becoming too lengthy and unmanageable.
8) Your submission will only consist of your Jupyter notebook. Your notebook should be run before saving it so that all analysis and visualizations are visible upon opening it. Points will be deducted if your notebook needs to be run before it is graded. Use the *P4LastNameFirstInitial.ipynb* format when naming your notebook.

PROJECT SUGGESTIONS

1) Data cleaning, munging and wrangling can be tricky. Start early to allow enough time to work through it.
2) Don't be afraid to ask for help. I am happy to answer general "how to" or "should I?" questions in class but suggest coming to office hours if you want help you to get your code to work.
3) Periodically run your notebook again from top to bottom. When you go back and make changes, these don't always trickle down as your cells run based on the order in which the cells were executed, not the order in your notebook. If we happen to rerun your notebook, you don't want things to break!