Time Series Data Processing Assignment

Preparing Economic Indicators for Housing Price Forecasting

Sonu Yadav, Sagnik Halder

1 Introduction

You are working as a data engineer for a real estate analytics firm. The data science team needs you to prepare various economic datasets for a machine learning model that will forecast housing prices. The datasets you've been given come from different sources, have different formats, and are sampled at different frequencies.

Your job is to process and unify these datasets into a format suitable for time series modeling.

2 The Datasets

Dataset Folder: All CSV files can be downloaded from:

`https://drive.google.com/drive/folders/1iPS1XHKvHilTJmBkWUqTH3M7Db_WcIie?usp=sharing`

You have been provided with the following CSV files. Each contains economic indicators that influence the housing market in different ways:

CSV File Name	**Contains**	**Housing Market Influence**
`Home Prices fred stlouisfed.csv`	S&P/Case-Shiller U.S. National Home Price Index	**Target Variable** - The price we aim to predict
`USpopulationpermonth multpl.csv`	U.S. population figures (monthly) with values stored as strings like "331.5 million"	**Positive** - Higher population increases housing demand
`Inflation Rates thebalance.csv`	Annual inflation rate percentages	**Complex** - Drives nominal prices up but affects affordability and purchasing power
`gross_domestic_product_usafacts.csv`	Annual GDP in billions of dollars	**Positive** - Economic growth increases income and housing demand
`Mortgage interest rates FreddieMac.csv`	30-year fixed mortgage rates (weekly)	**Negative** - Higher rates reduce affordability and cool demand

CSV File Name	Contains	Housing Market Influence
`USA Rental Rates fred stlouisfed.csv`	Consumer Price Index for rent of primary residence	**Positive** - High rents make buying more attractive, driving prices up
`U.S. Housing Starts investing.csv`	Number of new residential construction projects started, stored as strings like "1.5M"	**Negative** - Increased supply puts downward pressure on prices
`foreclosures corelogic.csv`	Annual number of completed foreclosures	**Negative** - Distressed sales depress market prices

Important Note

Not all CSV files need to be used in your final dataset. Part of your task is to select appropriate features and justify your choices.

3 Question 1: Working with String-Formatted Numerical Data

3.1 The Challenge

Two of your datasets contain numerical values that have been formatted as strings for human readability. You need to convert these to proper numerical formats before they can be used in any analysis or modeling.

3.2 Your Datasets

- `USpopulationpermonth multpl.csv`
  - Values formatted as: "331.5 million", "250.1 million", etc.
  - Monthly frequency
- `U.S. Housing Starts investing.csv`
  - Values formatted as: "1.5M", "1.2M", etc.
  - Monthly frequency

3.3 Your Task

Objective

Clean both datasets and convert string-formatted values to proper numerical formats suitable for analysis.

Things to consider:

- How will you strip the text suffixes while preserving the numerical values?
- What should the final units be? (Actual counts vs millions)
- How will you handle the date columns?
- What if there are missing or invalid values?
- One dataset is in reverse chronological order - how does this affect your processing?

Deliverable: Two cleaned CSV files with proper numerical data types and consistent date formats.

4 Question 2: Temporal Alignment Through Re-sampling

4.1 The Challenge

Your datasets are sampled at different frequencies:

- Some are **weekly** (e.g., mortgage rates)
- Some are **monthly** (e.g., population, rent)
- Some are **annual** (e.g., inflation, GDP, foreclosures)

For time series modeling, all features must be at the same temporal frequency.

4.2 Your Datasets

- `Inflation Rates thebalance.csv` - Annual data
- `gross_domestic_product_usafacts.csv` - Annual data
- `foreclosures corelogic.csv` - Annual data
- `Mortgage interest rates FreddieMac.csv` - Weekly data

4.3 Your Task

Objective

Resample all datasets to a consistent monthly frequency using appropriate interpolation techniques.

Things to consider:

- For upsampling (annual → monthly): Should you use linear, spline, or other interpolation? Why?
- For downsampling (weekly → monthly): Should you take the mean, max, last value, or something else?
- How do you ensure your interpolated values are realistic and don't introduce artifacts?
- What assumptions are you making when you interpolate data?
- How do you validate that your resampling makes sense?

Deliverable: Resampled CSV files at monthly frequency with justification for your interpolation choices.

5 Question 3: Dataset Integration and Feature Engineering

5.1 The Challenge

Now that all datasets are cleaned and at monthly frequency, you need to merge them into a single unified dataset. However, this isn't just a simple merge - you need to think carefully about temporal alignment, missing data, and how the final structure will be used.

5.2 Your Datasets

All processed CSVs from Questions 1 and 2, plus:

- `Home Prices fred stlouisfed.csv` (your target variable)
- `USA Rental Rates fred stlouisfed.csv`

5.3 Your Task

Objective

Create a single, unified dataset that contains all features aligned by date, properly handles missing values, and is structured appropriately for time series modeling.

Things to consider:

- What column should you merge on?
- What type of join should you use? (inner, outer, left, right)
- How will you handle rows where dates don't align perfectly?
- What should you do about missing values after merging?
- Should all features be on the same scale? If so, how?
- How do you ensure the target variable (House Prices) is properly positioned?
- What date range should your final dataset cover?

Deliverable: One unified CSV file with all features aligned by date, plus a brief explanation of your merging strategy and any decisions you made about missing data or feature selection.

6 What You're Preparing This Data For

After you complete this assignment, the cleaned and unified dataset will be used by the data science team to train machine learning models for housing price forecasting.

6.1 Why Your Data Processing Matters

The quality of your data processing directly impacts the success of any model built on top of it:

1. **Temporal consistency**: Features must be sampled at the same frequency so the model can learn proper relationships between variables over time
2. **Alignment**: Dates must align perfectly across features, otherwise the model will learn from incorrect feature combinations
3. **Data quality**: String formats, wrong units, or parsing errors will cause training failures or produce unreliable predictions
4. **Missing data**: Gaps in the time series can confuse sequential models and reduce prediction accuracy
5. **Proper formatting**: All data must be in numeric format with consistent units for mathematical operations and scaling

6.2 Requirements for the Final Dataset

Your processed dataset must be "model-ready":

- All features at monthly frequency
- No missing values
- All numeric data types (no strings)
- Features aligned by date
- Proper temporal ordering (oldest to newest)
- Consistent units across all measurements

The data science team will handle the subsequent steps of normalization, train-test splitting, and model architecture design.

Quality data processing is the foundation of successful machine learning.
Your work here enables the data science team to focus on model development.