U.S. and the living-cost expenses. Assessing the ability of households to afford house prices.

1. Introduction

1.1. Background: What is happening with the U.S. residents?

According to the latest research[1], for the last 4 years, the U.S. residents have been suffering from a cost-of-living crisis, leaving Americans drowned in pessimism. The "cost of living crisis" occurs when the expenses necessary for basic essentials, such as housing, food, healthcare rise faster than people's income levels. This rise in product pricing is happening due to the American dollar diminishing its value. In other terms, the current situation suggests that it takes a larger amount of money to buy the same item as before the price escalations. Due to this inflation, American families have lost an average of 4.3k dollars in annual purchasing power.

1.2. Reasoning behind topic selection

Due to the increased prices in these latest years, the U.S. households are facing difficulties in maintaining a good quality of life. An increase of living expenses could possibly affect people in various forms, considering low investment rates, and diminished housing affordability. The current studies[2] show that many people struggle to keep a roof over their heads, and many households are unable to purchase a home. In this regard, this is a sensitive topic that needs to be addressed; furthermore, it needs to be analyzed if the latest living costs of U.S. residents are affecting the people's ability to afford a place.

1.3. Hypothesis in interest of the reader

By economic theory, it is suggested that if the price of a product remains high for a very long time, something is affecting the supply-demand balance. This project suggests that there is a strong connection between the essential living expenses of the U.S. households and their ability to afford a house. To investigate this, this project aims to answer the question:

How do essential living expenses impact housing affordability for households across the U.S.?
To answer this question, 2 main data sources with recent data are considered.

2. Data Sources

Both datasets are sourced from Kaggle and are under the same standard open-data license. Creative Commons Zero v1.0 Universal (CC0 1.0). The license disclaims liability to freely copy, modify, and use without permission, in compliance with applicable law. *I plan to respect the license without violating the defamation regulations, and I will copy, use and transform the datasets under the license in a responsible manner, without taking for granted the openness and accessibility it offers.*

2.1. Dataset 1

The dataset, currently available on Kaggle, is derived from the Family Budget Calculator developed by the Economic Policy Institute (EPI).

Title	US Cost of Living Dataset (1877 Counties)
Metadata URL	Kaggle Dataset Page
Data URL	U.S. Cost of Living Data (Kaggle)
License	CC0 1.0
Data Type	CSV

[1] The Heritage Foundation. Facing a Cost-of-Living Crisis? You're Not Alone. https://www.heritage.org/budget-and-spending/commentary/facing-cost-living-crisis-youre-not-alone
[2] Business and society. The Market Alone Can't Fix the U.S. Housing Crisis https://hbr.org/2024/09/the-market-alone-cant-fix-the-u-s-housing-crisis

Key Content	This dataset contains information about the cost of living in 1877 US counties in specific columns: state, metropolitan status, county, family member count, costs for housing, food, transportation, healthcare, other necessities, childcare, taxes, total cost, and median family income. The data is mostly cleaned and well-organized. However, it needs transformations of columns for readability concerns.
Relevance to the project	For the aim of this project, this dataset offers valuable insights about households' essential expenses and their income across all 50 U.S. states in the last year. It supports the exploration of the connection between house affordability and basic life costs. This data is critical for further data analysis during the project.

2.2. Dataset 2

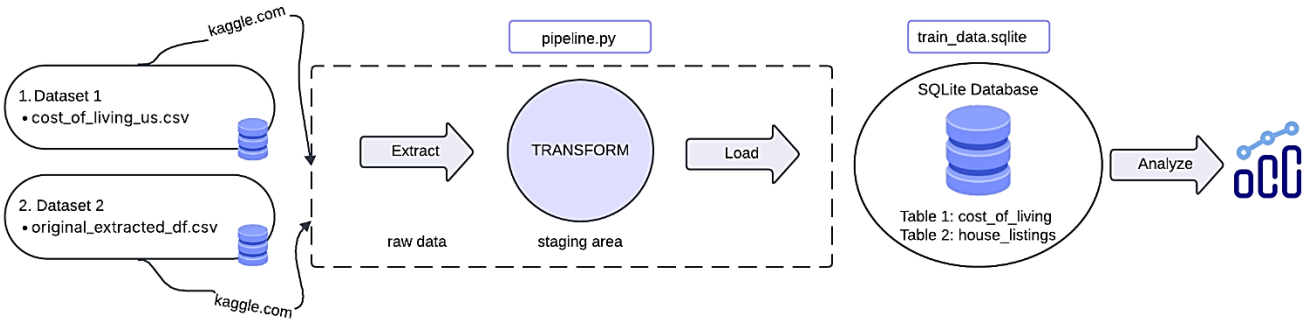
The dataset, currently available on Kaggle, is derived from Zillow Home Listings, a prominent online real estate marketplace in the U.S.

Title	United States House Listings: Zillow Extract 2023
Metadata URL	Kaggle Dataset Page
Data URL	U.S. House Listings Data (Kaggle)
License	CC0 1.0
Data Type	CSV
Key Content	This dataset provides insights into real estate trends by specific columns: State, City, Street, Zip-code, Bedroom, Bathroom, Area, Price Per Square Foot, Lot-Area, Market-Estimate in dollars, Rent-Estimate in dollars, Latitude, Longitude, Listed-Price in dollars. The data is not cleaned and has undefined rows. It needs to be properly cleaned and transformed.
Relevance to the project	This dataset offers gathered data on home listings, properties, and prices across several states. By focusing on the price column, it is possible to further analyze how much people would have to spend on housing in relation to what they earn. Consequently, it is possible to calculate the price-to-income ratio and, furthermore, the affordability index, which determines the answer to this project's main question.

3. ETL Pipeline

3.1. Overall view

The pipeline for this project is implemented in the Python language, and it uses libraries such as Kaggle API for dataset extraction, pandas for data transformation, and lastly SQLAlchemy for database integration. It extracts two datasets from Kaggle, transforms them in multiple ways, and then loads the restructured data into a SQLite database for further analysis.



Data is transformed for several reasons:

1. Reading and computation clarity.
2. To have data consistency for calculations.
3. To shift focus on the important data for calculations.
4. To ensure valid numerical data.

	Extract	Transform	Load
Function used	*download_kaggle_datasets(url, _path)*	*rename(columns={...}) drop(columns=[...]) fillna(value)*	*initialize_sqlite_db(db_name)*
Description	Loads datasets into pandas DataFrames after downloading them from Kaggle.	Renames columns, drops irrelevant columns, imputes missing values (using median).	Loads the cleaned data into SQLite tables.

3.2. Problems and solutions

Challenge	Solution
Automating dataset fetching	I implemented a custom function for Kaggle API, URL validation, and file processing.
Handling missing values	I applied median imputation for critical columns, and I dropped rows with all NaN values.
Managing errors during ETL	I added error checks and logging for invalid inputs.

Note: For detailed steps on setting up and running the pipeline, please refer to the pipeline.sh script on the ./project folder [3].

3.3. Meta-quality measures

Measure	Description
Dataset Validation	Checking the format of the URL and ensuring that only valid datasets are processed. Otherwise, raise an error if invalid. *raise ValueError("Invalid URL format.")*
Error Handling	Providing detailed feedback about errors during data processing.
Database Management	Handling the conflict cases between the outdated SQLite database files and reinitializations. Ensuring updated loading of tables. *os.remove(db_path)*
Data Integrity	Removing incorrect data, especially NaN values to ensure the usability of data for analysis.

4. Result and limitations

4.1. Result

Within the pipeline, I applied transformations to ensure that data has high quality. The cleaned and well-structured data is then saved in an SQLite database named train_data.sqlite. This database consists of two tables, cost_of_living and house_listings, each of which contains relevant columns necessary for later analysis. I used SQLite as the storage format, since it integrates easily with python through libraries, it is open-source, and it is easy to set up.

4.2. Limitations

- Currently median family income is consistent within counties. This could possibly overlook the income variance within households and doesn't consider different socioeconomic groups.
- There is a presence of outliers in several columns of both tables, which can distort measures like median and lead to biased results.

4.3. Possible issues

- For further analysis I would have to consider the median income column, which has the same values for each county but different for states. I would have to aggregate the data at a state level using a statistical measure, which can potentially skew the result if my chosen measure doesn't represent the majority data. My approach can oversimplify the diversity in incomes within a state.
- When calculating home prices, I would need to use a statistical measure to compute a representative value for each state. This approach could lead to skewed results due to the unequal housing markets within states, which differ between urban and rural areas.

[3] https://github.com/xxhensila/made-template/blob/main/project/pipeline.sh