“The Americas”
Projects with relation to North-, Middle- or South-America

Project Context:
The project examines a critical aspect of the American economy: housing affordability. By analyzing data on average income and housing prices across U.S. ZIP codes, the study explores whether homeownership—a key component of the "American Dream"—is achievable for people at different income levels. The analysis spans regional disparities, identifying factors that contribute to affordability challenges and providing insights into improving equity in homeownership opportunities.

Main Question:
What is the correlation between average income and housing prices in the United States, and how many Americans, based on income levels, can realistically afford to own a home in different regions?

Data Sources:

Data Source 1: American House Prices

- Metadata URL: American House Prices Metadata
- Data URL: Download CSV
- Description: Contains housing price and demographic data for major cities across the U.S., helping identify regional variations in housing costs.

Data Source 2: Average Income Per ZIP Code USA

- Metadata URL: Average Income Dataset Metadata
- Data URL: Download CSV
- Description: Provides granular data on average incomes across ZIP codes, essential for analyzing housing affordability by income bracket.

Data Pipeline:
The data pipeline was implemented using Python, Pandas, SQLite, and Matplotlib to automate the process of extracting, transforming, and loading data. The following steps and challenges were encountered during the project:

Technologies Used:

- **Python** for scripting and data processing.
- **Pandas** for data manipulation and cleaning.
- **SQLite** for storing cleaned data.
- **Matplotlib** for visualizing results.
- **Jupyter Notebook** for documenting and running analyses interactively.

Overview of the Pipeline:

1. **Data Extraction**:
    - Loaded data from two sources:
        - A CSV file containing housing prices and demographics.
        - A CSV file containing income data by ZIP code.

2. **Data Cleaning and Transformation**:
    - **Merging Issues**:
        - Addressed formatting inconsistencies and aligned state-level data.
    - **Column Selection and Renaming**:
        - Selected relevant columns (e.g., housing prices, average income) and renamed them to maintain consistency (e.g., zip_code, price, income).
    - **Validation and Data Integrity Checks**:
        - Ensured extracted columns matched expected schema.
        - Dropped rows with missing or malformed data.
3. **Data Integration**:
    - Merged datasets on ZIP codes to create a unified dataset containing both income and housing price data.
    - Calculated new metrics, such as the income-to-housing price ratio and affordability thresholds (e.g., houses costing more than 3–4 times the annual income are deemed unaffordable).

4. **Data Storage**:
    - Stored the cleaned and transformed data in an SQLite database for efficient retrieval and querying.
    - The database tables created were:
        - housing_prices_usa for housing price data.
        - income_usa for income data.

Key Challenges and Solutions:

1. **Inconsistent Data Formats:**
    - Challenge: Income and housing price data used different formats and regional granularity.
    - Solution: Implemented preprocessing pipelines to normalize data formats and align ZIP codes.
2. **Missing or Invalid Data:**
    - Challenge: Missing entries in income and housing datasets.
    - Solution: Dropped rows with missing data and applied validation checks to ensure consistency.
3. **Data Integration Issues:**
    - Challenge: Merging datasets with different levels of regional granularity (e.g., ZIP codes vs. states).
    - Solution: Focused on aligning data at the ZIP code level, which provides the most granular insights.

Data Analysis:

Affordability Analysis

- Homes priced at 3–4 times the annual income are considered affordable.
- Calculated the percentage of ZIP codes where the median house
- price exceeds affordability benchmarks.

- Identified regions where income levels are most misaligned with housing prices.

Results and Limitations:

Output Format

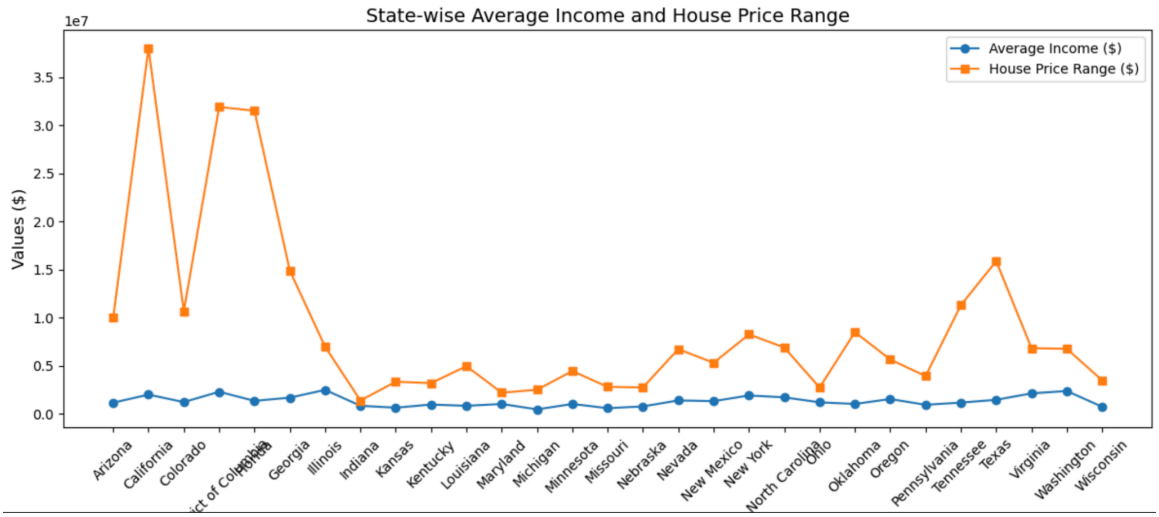
- Cleaned datasets stored in an SQLite database for efficient querying.
- Visualizations created using Matplotlib to compare regional trends and affordability metrics.

Limitations

1. Data Granularity:
    - ZIP code-level analysis may not capture nuances such as neighborhood-level disparities within cities.
2. Assumptions on Affordability:
    - Affordability benchmarks (3–4 times annual income) may not account for additional costs like taxes or maintenance.

Visualizations

- Bar Charts: Comparison of average incomes and housing prices by region.
- Scatter Plots: Correlation between income and housing prices.



Conclusion:

The study highlights significant regional disparities in housing affordability across the U.S., with income levels playing a crucial role in determining homeownership feasibility. These findings underscore the need for targeted policy interventions, such as affordable housing initiatives and income support programs, to address the growing gap between income and housing costs.