@larrycoder123

@etaaa

US County Project

This project aims to analyze the relationship between economic and social factors and household income across US counties. It involves the creation of a comprehensive dataset by integrating data from various sources, followed by aggregation and modification steps to ensure consistency. The dataset was then preprocessed for data visualization and used to develop a regression model predicting mean household income based on the available economic and social indicators.

Libraries:
- pandas
- matplotlib
- seaborn
- scikit-learn

Data Sources:
- US Census
- BLS
- Zillow
- ...

Methods:
- k-means imputation
- linear regression
- random forest
- ...

1. Data Aggregation

The dataset was constructed by collecting data from multiple sources, either by fetching it through APIs or importing CSV files. The collected data was then modified and merged using a common identifier, the "fips_code," to ensure consistency across different datasets. This data aggregation and preprocessing were carried out in the notebook data_import_modify.ipynb, which provides further details on the data sources, specific modifications, and a complete list of variables included in the final dataset.

2. Data Preprocessing

The merged dataset is further refined in project.ipynb to ensure its suitability for analysis. This includes grouping certain data points, removing columns that are not relevant for further analysis, and handling missing values. For counties with missing data, imputation is performed using the weighted average of the k-nearest neighbors, determined based on geographical distance, to maintain data integrity and consistency.

3. Visualization & Regression

With the final dataset, various plots were created to visualize the distribution of mean household income and its correlation with other variables. Mean household income serves as the target variable for the regression model. Both linear regression and a random forest regressor were implemented, and their performance was evaluated using relevant metrics to determine the better-performing model.

Architecture below

Architecture