Datasheet—Housing Price Predictions Model Data

Data Overview

- Original data *Housing Price Prediction Data* created by Muhammad Bin Imran.[1]
- Original data created to be explored by data science enthusiasts according to primary programmer.
- Used to create Housing Price Prediction model developed by Allie Craddock (alliec45@vt.edu), Esther Kim (estherkdy@vt.edu), and Haley Fore (haleyrhiann4@vt.edu), 2023, v1.

Motivation

- Purpose of the data is to help predict pricing for houses based on five of their different variables: size of the house, number of bedrooms, number of bathrooms, location of the neighborhood the house is in, and the year it was built in.

Composition

- Simulated dataset created with the software NumPy, which utilized a random sample command size set to 50,000.
- Inspiration for this dataset stemmed from online data collected from around the world.
- There are no external resources linked to this dataset, and it is all available within the dataset.
- Dataset contains a sample size of 50,000 sample points, each containing binary values of neighborhood type and numerical values for all other variables.

Collection Process

- Dataset was synthetically created by randomly generating values based on real-world trends.
- Data collection process was done without compensation.
- The timeframe during the collection process is unknown. Muhammad Bin. Imrad was the sole generator for the dataset.[1]

Processing/Labeling

- Data was prelabeled by original creator.
- Original data was created exceptionally clean, although not accurate.
- Only additional measures taken by model group were to remove all sample points that contained negative price values within the data.

Uses

- Dataset has been used to further understand housing prices based on their different variables.
- Repository to Housing Price Model is available by developers.[2]

Distribution

- Original dataset distributed through Kaggle.[1]
- Further cleaned data with removed negative values available via GitHub.[2]

Maintenance

- Maintenance for original data upkept by original creator.
- Model created with dataset hosted by Allie Craddock on GitHub.[2]
- Encountered errors will be acknowledged in future versions of the dataset within the same location via GitHub.

[1] Imran, M. *Housing Price Prediction Data*. (Version 1) [Data set]. https://www.kaggle.com/datasets/muhammadbinimran/housing-price-prediction-data/

[2] "Housing Price Prediction Model. (Version 1) [Data Set]. https://github.com/ENGL-3844-Writing-with-LLMs/group5-house-price-model