Dataset Proposal

Vikas Rayala, Manikanta Reddy Kallam

2023-04-06

Dataset Source and links:

Dataset description : House price prediction based on locality and neighbor cities.

File name: California_Houses.csv

Source : The original data (without the distance features) was initially featured in the paper: Pace, R. Kelley, and Ronald Barry. "Sparse spatial autoregressions." Statistics & Probability Letters 33.3 (1997): 291-297. They collected information on the variables using all the block groups in California from the 1990 Cens us

Link : https://www.kaggle.com/datasets/fedesoriano/california-housing-prices-data-extra-features?resource=download

Column level description and statistical analysis:

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
Median_House_Value	0	0	0	0	0	0	numeric	3842
Median_Income	0	0	0	0	0	0	numeric	12928
Median_Age	0	0	0	0	0	0	integer	52
Tot_Rooms	0	0	0	0	0	0	integer	5926
Tot_Bedrooms	0	0	0	0	0	0	integer	1928
Population	0	0	0	0	0	0	integer	3888
Households	0	0	0	0	0	0	integer	1815
Latitude	0	0	0	0	0	0	numeric	862
Longitude	0	0	0	0	0	0	numeric	844
Distance_to_coast	0	0	0	0	0	0	numeric	12590
Distance_to_LA	0	0	0	0	0	0	numeric	12590
Distance_to_SanDiego	0	0	0	0	0	0	numeric	12590
Distance_to_SanJose	0	0	0	0	0	0	numeric	12590
Distance_to_SanFrancisco	0	0	0	0	0	0	numeric	12590

Response Variable :

1. Median_House_Value : Median house value for households within a block (measured in US Dollars) [$]

Potential Predictor variables :

1. **Median_Income :**Median income for households within a block of houses (measured in tens of thousands of US Dollars) [10k$]
2. **Median_Age :** Median age of a house within a block; a lower number is a newer building [years]
3. **Total_Rooms :** Total number of rooms within a block
4. **Total_Bedrooms :** Total number of bedrooms within a block
5. **Population :** Total number of people residing within a block
6. **Households :** Total number of households, a group of people residing within a home unit, for a block
7. **Latitude :** A measure of how far north a house is; a higher value is farther north [°]
8. **Longitude :** A measure of how far west a house is; a higher value is farther west [°]
9. **Distance_to_coast :** Distance to the nearest coast point [m]
10. **Distance_to_Los Angeles :** Distance to the centre of Los Angeles [m]
11. **Distance_to_San Diego :** Distance to the centre of San Diego [m]
12. **Distance_to_SanJose :** Distance to the centre of San Jose [m]
13. **Distance_to_SanFrancisco :** Distance to the centre of San Francisco [m]

Conclusion:

From the above table, we can see there are no missing values for any of the variables. So There is no need of data preprocessing and we can move ahead with analysis and data modelling.