**Data Science: House Pricing Data & Income Qualification Data.**

Project 101

DESCRIPTION

**Background of Problem Statement :**

The US Census Bureau has published California Census Data which has 10 types of metrics such as the population, median income, median housing price, and so on for each block group in California. The dataset also serves as an input for project scoping and tries to specify the functional and nonfunctional requirements for it.

**Problem Objective :**

The project aims at building a model of housing prices to predict median house values in California using the provided dataset. This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics.

Districts or block groups are the smallest geographical units for which the US Census Bureau  
publishes sample data (a block group typically has a population of 600 to 3,000 people). There are 20,640 districts in the project dataset.

**Domain**: Finance and Housing

DESCRIPTION

Identify the level of income qualification needed for the families in Latin America.

**Problem Statement Scenario:**  
Many social programs have a hard time ensuring that the right people are given enough aid. It’s tricky when a program focuses on the poorest segment of the population. This segment of the population can’t provide the necessary income and expense records to prove that they qualify.

In Latin America, a popular method called Proxy Means Test (PMT) uses an algorithm to verify income qualification. With PMT, agencies use a model that considers a family’s observable household attributes like the material of their walls and ceiling or the assets found in their homes to  
classify them and predict their level of need.

While this is an improvement, accuracy remains a problem as the region’s population grows and poverty declines.

The Inter-American Development Bank (IDB)believes that new methods beyond traditional econometrics, based on a dataset of Costa Rican household characteristics, might help improve PMT’s performance.

Following actions should be performed:

* Understand the type of data.
* *Write a function* print\_various\_data\_info with input parameter as DataFrame. This function will *print various info of DataFrame (How many entries, how many columns, column name(s), how many nont null columns for each column, DataType of each column, Shape of data, Null count for each columns.*
* Print the first 10 rows. Print the last 10 rows,
* *Write a function print\_numerical\_columns with input parameter as DataFrame to print numerical columns (Int/Float). Return column list.*
* *Write a function* get\_int\_float\_dtype\_null\_column\_list

*with input parameter as DataFrame to print numerical columns (Int/Float) containing Null value. Return column list.*

* *Write a function* fill\_int\_float\_dtype\_null\_cells\_data\_with\_mean with parameters as DataFrame and column\_list *to fill Null cells of int\_or\_float column(s) with mean() value of the corresponding column.*
* *Checking if there any Duplicate rows*
* *Write a function* get\_object\_dtype\_column\_list with parametsr as DataFrame *to get the list of all Object Dtype Column(s). Return column list.*
* ***Note: This is the only place you have to use House Pricing Data file. Rest of the places you use Income Qualification data. >>>*** *Analyzing the value\_count of various types of 'ocean\_proximity' to see if it has limited set of data for encoding purpose*
* *Write a function* get\_category\_column\_list with parameters as DataFrame and object\_column\_list *to get Object Dtype categorical column(s) having <= 10 diff type of data. Return* category\_column\_listand category\_column\_value\_list. Second one is list of list.
* Write a function apply\_category\_column\_encoding with parameters as DataFrame, category\_column\_list, category\_column\_value\_list and create additional columns and drop the original column. Additional column will contain 1 or 0 depending upon the corresponding column and categorical value.
* Check if there are any biases in your dataset.
* Check whether all members of the house have the same poverty level.
* Check if there is a house without a family head.
* Set poverty level of the members and the head of the house within a family.
* Remove null value rows of the target variable.
* Marge the two DataSets