**Data Cleaning and Preparation Report**

Dataset Overview (Before preparation)

Rows: 20,640

Columns: 10

File Name: housing new.csv

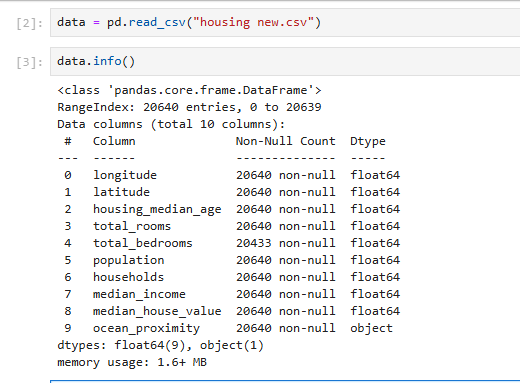
**Column Summary**

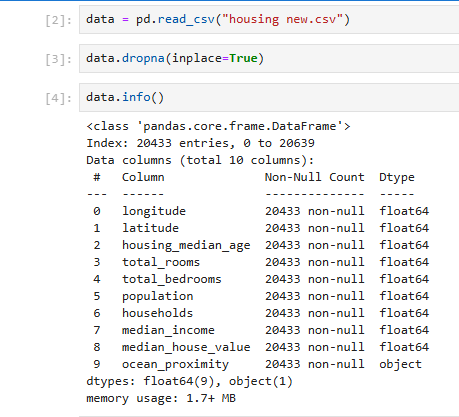
**1. longitude**: A measure of how far west a house is; a higher value is farther west  
  
**2. latitude**: A measure of how far north a house is; a higher value is farther north  
  
**3. Housing\_Median\_Age**: Median age of a house within a block; a lower number is a newer building  
  
**4. totalRooms**: Total number of rooms within a block  
  
**5. totalBedrooms**: Total number of bedrooms within a block  
  
**6. population**: Total number of people residing within a block  
  
**7. households**: Total number of households, a group of people residing within a home unit, for a block  
  
**8. medianIncome**: Median income for households within a block of houses (measured in tens of thousands of US Dollars)  
  
**9. medianHouseValue**: Median house value for households within a block (measured in US Dollars)  
  
**10. oceanProximity**: Location of the house w.r.t ocean/sea

**Missing Values**

Only one column contains missing data:

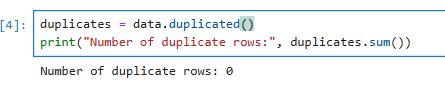
* total\_bedrooms: **207 missing values** (1% of the data)



* To take care of these missing values, the best way was the removal of all rows with missing values to leave us with **20,433** rows in each column.
* 

**Duplicate values**

The data set used had no duplicate values in it



**Data Types & Consistency**

* 9 numerical columns: Properly recognized as float64
* 1 categorical column (ocean\_proximity): Recognized as object
* Converted it into multiple **binary columns** (one for each category) as we only had 5 categories through one hot encoding.
* 