APPLIED DATA SCIENCE WITH PYTHON

FEATURE ENGINEERING REAL ESTATE ANALYTICS PROJECT

Explanation:

1. First we have inserted the csv file from the local storage using .read_csv() function and viewed the last and first 5 rows using head() and tail() function

2. Next we have seen the no of rows and no of columns using shape attribute

3. To find the null values from the data frame and its count we used the .isnull().sum() and resulted shows the null values from each column of the dataframe but here we could not filter and locate exact columns which has null values because the no of columns very high in the dataset.

4. Now to filter the columns which exact has the null values ,we have used the logic of for-loop and iterated each column for any null values and if found we push the column into the empty list with display the count and if the column does not have any null values then we skip the iteration using the continue - which is loop control statement

5. Next we have filtered the numerical columns from dataframe using for-loop logic and stored separate dataframe object

6. Next we have filtered the categorical columns from the dataframe by using subtracting the previously obtained column dataframe with the original data frame thereby getting the categorical columns and stored the same in new dataframe object

7. Now we have printed the new dataframe using newly obtained categorical and numerical data frame objects .

8. Next step we do the data cleaning using dropna() function where we drop the rows that has 30% missing values and we used the drop() function to drop the columns that has 90% of missing values because these type of missing values does not contribute much to the final output and it is also not useful while we fit the cleaned data frame to the model

9. Next Step we do the EDA process which means exploratory data analysis and test for the skewness of the data frame using the distribution plot(combination of histogram +density plot)

10. Next we visualize the correlation matrix with the help of a heatmap where Brighter colors represent high activity/high correlation areas,dull shades represent low correlation.

11. Next we generate the pairplot , countplot and swarmplot to get more insights about the variables in the dataframe and their relationship.

12. Next task is to perform the chi-squared test among any two variables to determine the hypothesis and if the p-value is less than 0.05 then we reject the null hypothesis orelse we accept the null hypothesis