Maddi Monclús - Python Project

I have been given by my client a table which contains data about houses. The database has 1460 records with data about 80 different variables, information such as Sale Price, number of rooms in the house, the area, whether they have a garage and when it was built, and many other data. I have been asked to complete an analysis of the data and to provide a report.

The following report is organised into two parts. The first one explains the steps taken to clean the data, providing arguments about the reason the steps taken have been followed. For the second part, a data analysis has been performed, which can be found onthe second page of the report.

1. CLEANING DATA

First, a Data Cleaning needs to be completed, to be able to analyse the data. In an organised way. The data type has been checked in case anything needs changing, but everything seems correct. No typographical errors have been found.

In regard to missing data, there are 6965 records without data, which is a large number. Completing a quick visual check of the DataFrame, it can be seen that there are certain variables with almost no data, such as Alley, PoolQC or Fence. For this reason, columns that have over 30% of missing data will be removed from the DataFrame. After this task, the amount of missing data has been reduced to 868, where 348 are numerical missing data and 520 categorical missing data.

In the following steps of this part of the report, it has been explained, first, how the numerical missing data has been cleaned and after, the categorical missing data.

1.1. Numerical Data

A new DataFrame of only numerical data will be created to deal with this type of missing data. In order to complete all numerical missing data, the k-nearest neighbors (KNN) algorithm will be used. This algorithm uses proximity, it groups the data according to their similarity and it is used afterwards to make predictions and to fill missing data that has been found in the table. The k=5 refers to the number of nearest neighbors.

As when completing the transformation and filling in the missing data, the new dataframe has lost the column names. The column names of the numerical dataframe will be copied to the new one without missing data. To complete this task, the names of the variables/columns will be taken.

1.2. Categorical Data

Once all the numerical missing data task has been completed, the cleaning will be focused on the categorical data. Hence, a new DataFrame will be created only with the categorical data.

To fill the missing data of this DataFrame, the missing data will be filled by the most repeated and frequent value of each variable.

1.3. All Data

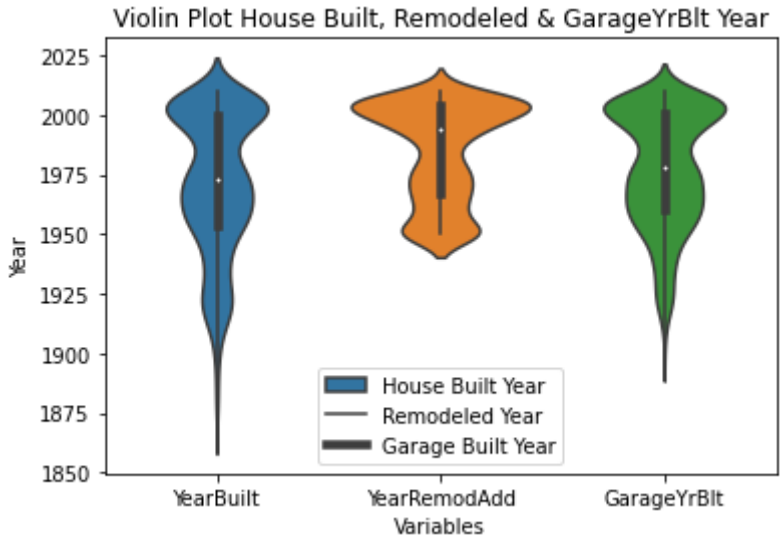
A new DataFrame will be generated using the categorical data dataframe and the numerical data dataframe created in the previous steps. Finally, after double checking it, it can be ensured that there is no missing data. This DataFrame will be used afterwards to be analysed.

2. DATA ANALYSIS

2.1. Violin Plot - year the houses were built & the year houses were remodeled

The Violin Plot that can be found below, shows a distribution of numeric data using density curves of the year the house were built, remodeled and the garage built. It needs to be mentioned that the data does not correlate regarding the same houses, but data is generic about houses.

The width of each curve corresponds with the approximate frequency of data points in each region. It can be observed that the number of houses built started increasing steadly after the year 1900. It can also be seen that there have been 3 main periods when there has been an increasing in the number of houses built. Also, the largest number of houses were built between 1985 & 2010.

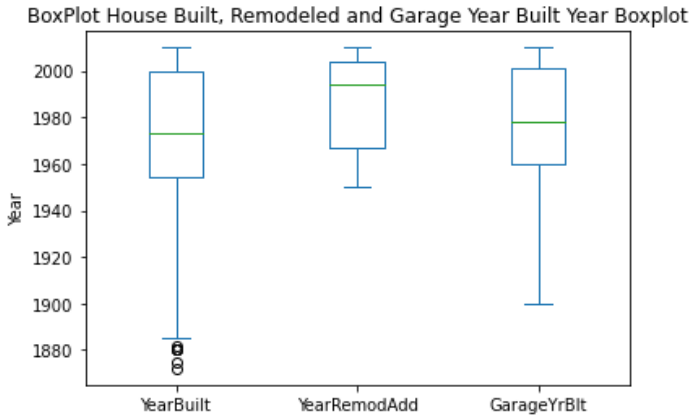


Besides, until 1950 there was no data with regard to the year the houses were remodeled, as well as the fact that most of the houses were remodeled between 1985 & 2010, following the pattern found previously forthe largest number of houses when they were constructed.

Finally, regarding the years the garages were built, it shows similar data comparing to when the houses were built and the periods.

To summarise, the highest periods in the three variables have been very similar. The data presented in the table below shows the statistical data of the three variables also shown in the graph above. For a visual representation of the distribution of the data, please check below a BoxPlot graph of the table.

Index	YearBuilt	arRemodAc	GarageYrBlt
count	1460	1460	1460
mean	1971.27	1984.87	1977.28
std	30.2029	20.6454	24.7495
min	1872	1950	1900
25%	1954	1967	1960
50%	1973	1994	1978
75%	2000	2004	2001
max	2010	2010	2010



2.2. BoxPlot - Sale Price of the houses

The BoxPlot graph shows how the houses' sale price data is distributed. The 50% of the prices, the box on the graph, are found between 129 975 and 214 000, being the mean 180 921.

The dots seen in the graph are the detected outliers. An outlier is an observation that lies at an abnormal distance from other values in the sample used, a value that lies outside (much larger or smaller) than most of the other values in a set of data. This group that consists of outliers will be taken for further analysis.

To obtain all the data from the cleaned DataFrame from only the SalePrice outliers rows, a new DataFrame has been created using the following join function, as per when the index of the outliers dataframe appears in the cleaned DataFrame.

** A further analysis of the outliers will be completed shortly*

2.3. PairPlot

At this point, a new DataFrame is generated only with data of the SalePrice outliers and their LotArea data. To check whether there is a correlation, a pairplot has been created. No correlation has been found between the two variables.

**A further analysis will be completed shortly for a better understanding*