Cleaning Data and EDA

Assignment 1

David Lizama

Abstract

This is a database of 5000 observations and 16 variables. The main objective of this project is to clean the data and do an exploratory analysis to find the most important elements that impact on the price of the houses.

Raw Data:

The original database is composed of these columns:

- MLS: ID of the house.
- sold_price: Price of the house
- zipcode: zip code of the house
- longitude: longitude of the location of the house
- latitude: latitude of the location of the house
- lot_acres:number of acres in the house.
- taxes: Amount of money is paid to the government for the house.
- year_built: Year that house was built.
- bedrooms: Number of rooms.
- bathrooms: Number of bathrooms.
- sqrt_ft: Square footage of the house.
- garage:number of garages.
- kitchen features: list of appliances in the kitchen.
- fireplaces:number of fireplaces.
- floor covering: type of floor covering.
- HOA: number of people in HOA organization.

In this case, MLS and zipcode are not considered for exploratory purposes. The first one because it is only an identificator of the house, and the second one because it is quite similar to the latitude and longitude of the house.

Cleaning Data:

Sold_price: It is the target variable and not cleaning was done.

longitude: Some of the observations had numbers separated by commas, so these records had to be changed. Also, missing values were filled in with the median of the distribution; it was not considered to use the mean due to asymmetric distribution, and there was no mode. Finally, outliers were treated to set a range between -109 to

-113 because some records had locations in the middle of the ocean; dot positions were set equal in all records to be sure the longitude of the house is in Houston. The type of the column was changed from object to float.

latitude: Missing values were filled in with the median of the distribution; it was not considered to use the mean due to asymmetric distribution, and there was no mode. Finally, outliers were treated to set a range between 30 to 35 because some records had locations in the middle of the ocean; dot positions were set equal in all records to be sure the latitude of the house is in Houston. Finally, the type of the column was changed from object to float.

lot_acres: Ten missing values were found and were filled in with 3. The mean, median and mode are equal to 3.

taxes: This column did not be treated.

year_built: There were some observations which had a record of zero. It is not possible, we assume zero is the number of years the house has been built, so we will replace zeros with last year recorded in the database.

bedrooms: This column did not be treated.

bathrooms: Six missing values were found and were filled in with the mode. The mean is 3.82 (4 bathrooms) and the mode is 3. In this case, we opt for the mode due to bias distribution to the left.

sqrt_ft: 56 observations with the term ‘None’ were changed to 3512 which is the median of the distribution. We choose the median due to the asymmetric distribution.

garage: 7 missing values were found and were filled in with 3. The mean is 3.82 (4 bathrooms) and the mode is 3. In this case, we opt for the mode due to bias distribution to the left.

kitchen features: There are a lot of features referring to the floor of the house. The most popular features were selected and new columns were added to fill in with ones in case the observation of the column had a particular keyword. These are the columns that were added and the number of occurrences:

● KF_Dishwasher	4857
● KF_GarbageDisposal	4520
● KF_Refrigerator	4234
● KF_DoubleSink	1164
● KF_Microwave	3625
● KF_Oven	3977
● KF_Compactor	432

KF_Freezer	395
KF_ElectricRange	401
KF_Island	1252
KF_GasRange	1307
KF_Countertops	1482
KF_Desk	327

.

fireplaces: There were 25 observations with missing data. The majority of the elements are closed to the mean and median (2 fireplaces), so we fill missing values with 2.

floor covering: There are a lot of features referring to the floor of the house. The most popular features were selected and new columns were added to fill in with ones in case the observation of the column had a particular keyword. These are the columns that were added and the number of occurrences:

FC_Stone	1499
FC_Ceramic	2527
FC_Laminate	86
FC_Wood	1248
FC_Carpet	3509
FC_Concrete	756
FC_MexicanTile	660

HOA: Some observations were separated by commas, so it should be replaced by dots. Additionally, there were 562 missing values and they should be replaced by the mode of the distribution because the distribution has a bias to the left, and the majority of the observations do not have a HOA. There were bad records of the number of people in the association, for example: 4.67 people, in this case, all the observations were changed to an integer number.

Data analysis

- The most important information to predict house prices are square footage, year of constructions, bathrooms, lot_acres, and fireplaces.
- There is not a positive association between longitude and latitude, and price of houses, but we can split specific areas in which we can find high value houses.
- Low association of house prices and the rest of variables. For example, features of kitchen and floor covering do not seem to be relevant.