Evan Carter
03/06/2022
Milestone 1

Report: Preliminary Inspection and Exploratory Data Analysis

Context

One of the most fundamental components of any economy is the housing market. Real estate prices, in addition to serving as a proxy for measuring the health of the economy, are of great interest to both homebuyers and homeowners when determining the best time to buy or sell a home. Numerous factors that can influence the listing price of a home lead to a market that is both sensitive and constantly in flux. Because of this, being able to make reliable predictions towards the optimal time for buying or selling a property would be great interest to any party involved in the investment of real estate.

Objective

Our objective in this project is to generate a data-driven model that can predict the housing prices of a locality based on various features of the homes and identify the most important features to consider in predicting said prices. To do this, we will make use of observational data introduced further on in this report to build a linear regression model that will be able to predict the price of a home based on features informed by the data.

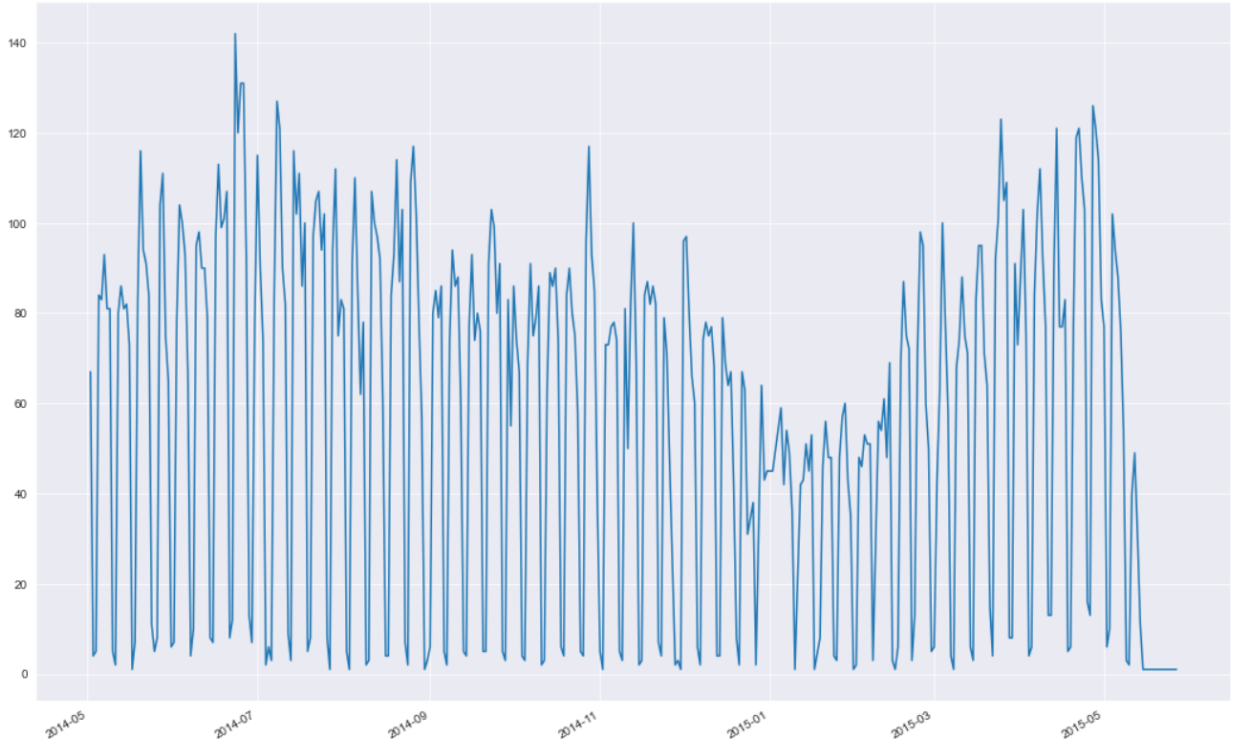
Data Dictionary

The dataset used in this project consists of 23 total features, including the sale price of a property as the target variable. The details of each of the features is provided below:

- cid – an identifier for a particular property.
- dayhours – the date that a house was sold
- price – the sale price of a house in USD (target variable.)
- room_bed – the number of bedrooms listed for the property.
- room_bath – the number of bathrooms listed for the property.
- living_measure – square footage of the house.
- lot_measure – square footage of the lot where the property is located.
- ceil – total floors listed for the property.
- coast – Set to ‘1’ if the property features a waterfront view or ‘0’ otherwise.
- sight – property has been viewed.
- condition – perceived condition of the property, rated out of a maximum of 5.
- quality – grade given to the property based on condition.
- ceil_measure – square footage of the house apart from any basement.

- basement_measure – square footage of the basement, if applicable.
- yr_built – year that the property was built.
- yr_renovated – year when the home was renovated, if applicable.
- zipcode – ZIP code for the property address.
- lat – latitude coordinates for the property.
- long – longitude coordinates for the property.
- living_measure15 – square footage of the living room in 2015. Could reflect renovations done on the property and may or may not affect size of the lot.
- lot_measure15 – square footage of the property lot in 2015. Changes could be due to renovations done on the property.
- furnished – set to '1' if the property was considered 'furnished' in the sale, '0' otherwise.
- total_area – measure of both living room and lot.

Preliminary Inspection of the Data

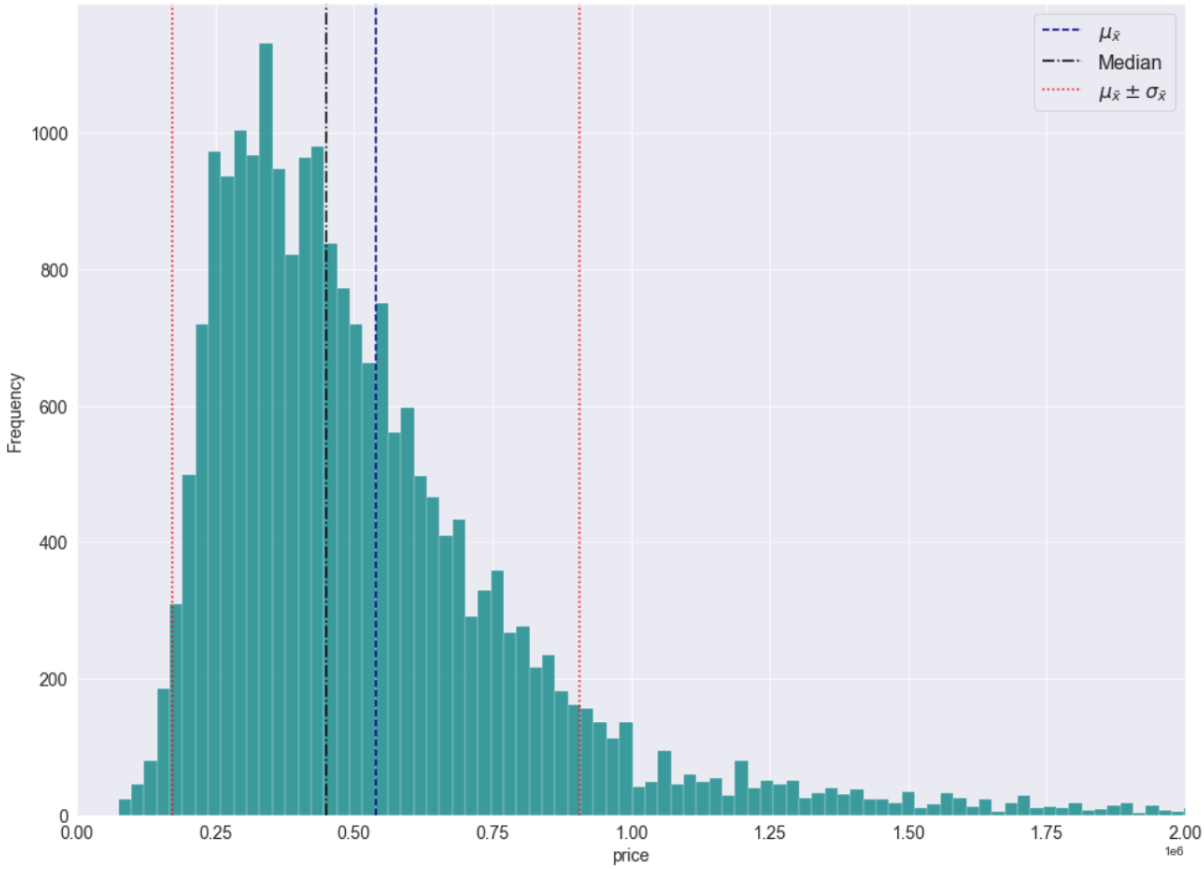


Simple Time-Series Plot of Total Home sales from 05/2014 to 05/2015

The data consists of 2613 observations on houses sold over a 1-year period from May of 2014 to May of 2015. Approximately 177 properties are included more than once in the data but upon inspection it was confirmed that the properties were sold more than once within the time period and properly included for different sales. Several of the observations contain missing values or incomplete data but no feature present in the data set is missing more than 0.5% of total values. Some of the features present in the data contain information that is highly

redundant, including different features regarding square footage and may not be included for use in building the final model.

Univariate Data Analysis



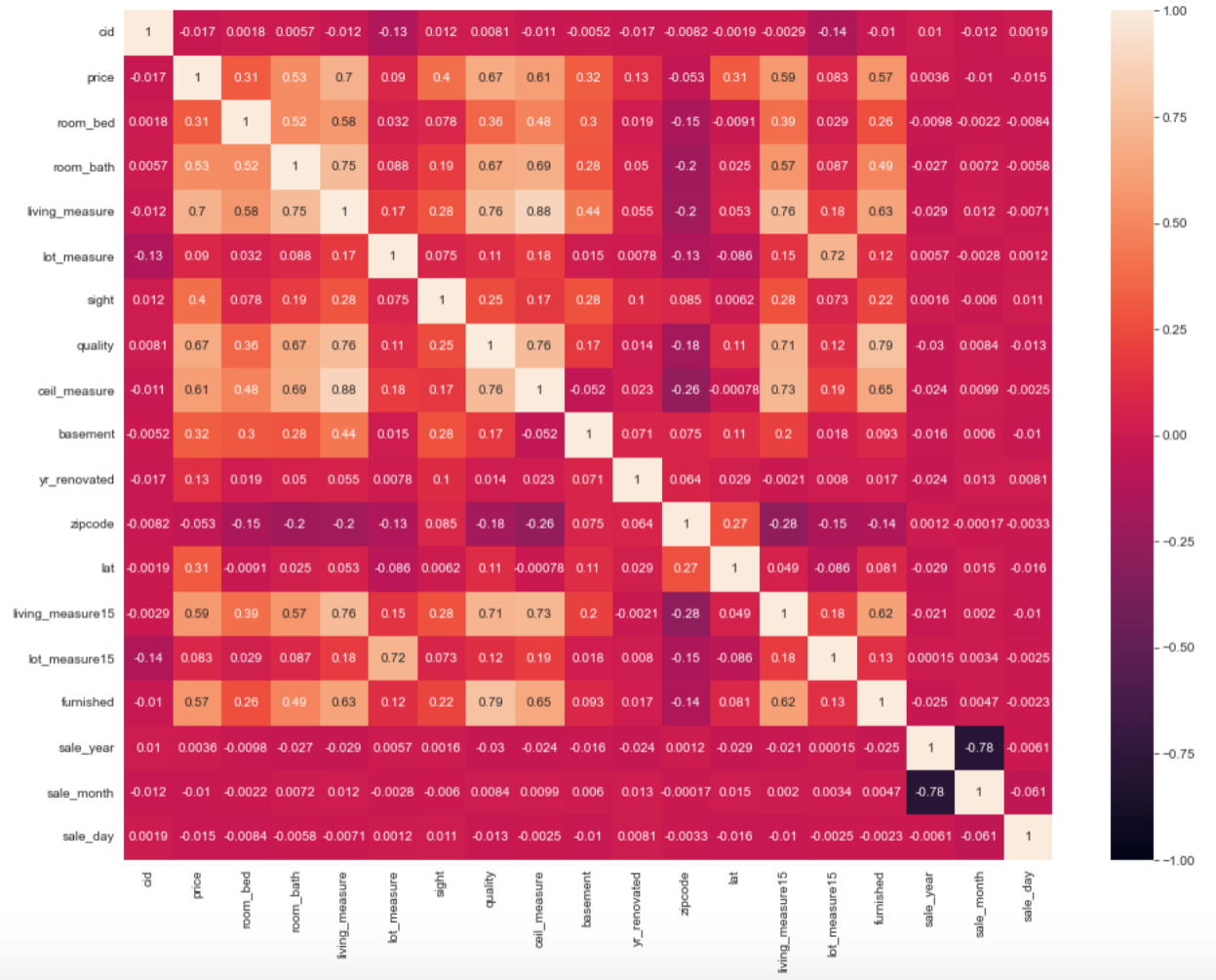
Distribution of sale prices included in the data, which is the target variable our model will predict.

Our initial observations and insights from performing univariate exploratory analysis are summarized as follows:

- The data contains observations for houses sold within a 1-year period from May of 2014 to May of 2015.
- About 2/3 of the sales data that we have are for homes sold in 2014.
- The houses were built between 1900 and 2015, with the majority of the houses built after the year 2000.
- While some of the homes have undergone renovations, the majority (~95%) have not been renovated.
- Most of the homes included in our observation are between one to two-story homes that are not located along a coast.
- Most of the homes feature between 2 - 5 bedrooms, 1 - 2.5 bathrooms and do not feature a basement.

- The majority of the homes were not sold as furnished but about one-fifth of the homes (19.6%) were sold as furnished.
- Interestingly, most of the houses were recorded as "not viewed" in the dataset. Given that most houses are not sold sight-unseen, this leads us to believe that the meaning of this feature is that the home was not viewed by the data collectors. Further clarification is needed on this feature.

Bivariate Data Analysis



Heatmap illustrating the correlation matrix between numerical features.

The results of our initial insights from multivariate analysis are provided here:

- Highest correlations shown in the matrix are due to features that would be expected to show high multicollinearity. Examples include the number of floors a property has and its area in square footage, the size of the lot and the sale price, whether a home is furnished and the quality rating, etc.

- The feature that shows the highest correlation with sale price is the square footage of the home, showing a 70% correlation. There is a potential for multicollinearity between this feature and the target as larger homes correlate with higher prices.
- Other strong correlations with price include rated quality out of 5 (67%), number of floors (61%), measure of the living room in 2015 (59%), whether the home is furnished (57%,) and the number of bathrooms (53%).
- Interestingly, the number of bedrooms does not show a particularly strong correlation with sale price (31%), even though there is a positive association with the number of bathrooms in a home.
- The most significant challenge we will face in building a successful predictive model will be how to mitigate the effects of multicollinearity between features that have a strong association.

Conclusion: Next Steps

After concluding with our initial exploratory data analysis, our next steps in the project will be as follows:

- More rigorous preprocessing of the data including missing value treatment, treatment of outliers or transformation of features that feature a high dynamic range of values such as square footage of property lots.
- Multivariate analysis of the geographical data including coordinates and ZIP code data, we are still working to determine how to properly include this information in our analysis.
- The total_area feature was experiencing an issue when we were attempting to visualize it, we will need to determine what the cause of the issue was so that we can resolve it.
- Features that exhibit high multicollinearity or redundant information will need to be carefully considered for inclusion in the model. We will calculate variance inflation factors for each parameter to help us arrive at a course of action for treating these variables.