Robin Singh

Assignment 1

Section 1: Sample Definition

* Remove houses with square footage above 4000 because of the scarceness of the data after that point may not allow for the best analysis of data.
* Remove houses with quality index below 5 and quality index below 2, as these houses are irregular in status after analyzing the range of values both features could take (with a quality index high of 35 and an overall quality high of 10). These houses seem to be extremely poor in quality and can hence cause a lot of variability in analysis. We should focus on houses with more moderate levels of quality and condition.
* Dwellings with MS Zoning classified as Commercial and Industrial should also be eliminated from the analysis and regression to avoid confusion with houses in residential areas like the rest. Commercial and industrial properties are significantly different than residential properties, resulting in a different pricing structure, which may be a point or error in conducting analysis.
* Lot shape IR3 also only have a few amounts of examples in this dataset, and it may not be enough to do a proper analysis on, and hence should be dropped.
* Sale conditions that are abnormal, Alloca, and Partial should not be considered as the sale prices may vary due to these circumstances, which may cause errors during analysis.
* Houses without typical Functionality should be removed.
* Lot area above 100,000 is a lot higher than the mean of lot value, and anything above that value seem to be outliers. Hence, houses with lot sizes should be removed, primarily because of the scarcity of data available above that lot size.
* We should also remove entries with N/A in the values being analyzed, so that the analysis may take the values it has into account, and it can prevent errors in the analysis down the line.

Section 2: Data Quality Check

• 20 Analyzed features:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Total Floor SF | House Style | Lotshape | House Age | MS SubClass |
| Lot Area | Lot Frontage | Overall Qual | Overall Cond | Year built |
| Exterior Quality | Exterior Condition | Misc Val | 1st Flr SF | 2nd Flr SF |
| BedroomAbvGr | TotRmsAbvGrd | Foundation | GrLivArea | YearRemodel |

* After viewing a histogram of TotalFloorSF, I can see there is a scarce amount of data after about 4000 square feet, so those above that value should be removed. This also has a significant impact on sale price, in my opinion, because house price tends to increase as house size does.
* The house style is in correlation with house size, so it may impact the price in a significant way as well.
* Lotshape also is related to overall property size. The more property, or higher the lot size, the higher the price should be, holding all else constant.
* House age could determine house quality, as newer houses may have more features and may last longer, as well as be in less need of repairs. The better the condition of the house, the higher its price should be. This needed to be edited due to the lack of abundant data for 'IR3.' It would be better to eliminate houses with that entry due to the lack of data.
* Subclass is also related to house style and house size, as 2 story houses are bigger and have more living area than a 1 story house, and hence could be more expensive holding other variables constant. This feature holds a lot of weight, so should be analyzed in relation to the other features.
* Lot area is another feature that contributes to the size of the property.
* Lot frontage also contributes to size via lot area, and seems to be a heavy weighted feature. It can also determine if the property is in a more isolated area or is connected to the street. Properties in the two different areas could vary in price significantly.
* Overall Conditions and Overall Quality both go hand in hand, and state how well the property is. The higher the value the higher the price typically is, hence making it a high weighted feature.
* Year built is related to the age, which may distinguish quality. SOme individuals may also prefer older homes due to material choice, which could contribute heavily to hous price.
* Exterior quality and conditions also contribute to how much repair the house needs. If the house is in need of more repairs and both values are low, their prices may be lower as well, making it a heavy weighted feature for analysis.
* Miscellaneous Feature Value also adds to property value significantly. Houses with a pool, shed, basketball court, etc. may be a more desirable property and could hence increase the price.
* 1st floor square feet and 2nd floor square feet contribute to the size of the property. It could also distinguish the house subclass, making it a high weighted feature.
* Bedrooms and total rooms above ground also can determine the size of the property and living space. It could also determine if larger or multiple families could live in the house. Distinguishing between the two could cause variances in price, making it worth including in analysis.
* Foundation could determine the material the houses were made of and could help determine the longevity of the property. This needs to be made into numerical values, however.
* GR Living Area contributes to size and how much space families have to live in the residence, contributing to price heavily in the end.
* The year the property was remodeled could significantly impact the style, features, and durability/quality/condition of the house, making it a heavily weighted feature.

All of these features are what I thought would contribute the most to determining the sale price of the house and are heavily weighted features. Some needed to be edited because of lack of data surrounding the various entries, which may cause greater errors in predictions.

Section 3: Initial Exploratory Data Analysis

* Out of the twenty features checked above, these ten seemed to be the more important features, as they were more numeric and easily analyzable. After graphing the variables, I could see the distribution of the variables. The ones with relatively normal distributions could be analyzed well, while some may need more preprocessing before it can be included in the model, such as Second floor square footage, as a lot of it is skewed to the right. This may be due to the abundance of homes that are one story, resulting in 0 values for 2nd floor square footage. Upon analysis this should be replaced, along with 1st floor square footage to just include all of the living area in the property, by including the feature 'GrLivArea.'

Section 4: Exploratory Data Analysis for Modeling

* I chose the three variables I thought would have the most linear relationship to sale price, as well as a categorical feature: TotalFloorSf, QualityIndex, and Lotshape (being the categorical). TotalFloorSF and QualityIndex seem to have quite the liner relationship to sale price, as they both seem to increase with each other for the most part. However, it is not extremely linear, showing the importance the other features may have on the sale price as well. Lotshape does not seem to have much of a relationship to sale price, as according to the boxplot, it is all over the place. However, this shows how many outliers are in this feature, and how it may need more preprocessing before being included in the model. This allows us to get an insight as to which features are heavily weighted and appear necessary for proper and more accurate analysis while which others need to be edited, omitted, or replaces in the model.

Section 5: Summary/Conclusions

* This assignment allowed me to explore the data and clean it up to find features that may be the most beneficial to creating a model, or at least those that may have the greatest impact. It also allowed me to warm-up on my R and R Studio skills to get ready for the assignments to come, which may involve modeling and more statistical analysis. It prepared me to read the data and visualize which ones may be well implemented if it were to be plugged into a model, and which ones need editing, should be combined with others for better characterization of a feature, and omitted as they may not contribute to the model in a positive way. The various visual aspects conducted in this assignment helped further develop the skillet of analyzing data and exploring ways to create a more accurate model.