# ANALYSIS ON THE AMES HOUSING DATASET

Robin Singh

MSDS 411 Summer 2021

## Table of Contents:

## Introduction:

Different properties consist of multiple different features that correlate with other features within certain areas or price groups. The Ames Housing dataset is a dataset that consists of 2,930 housing records within the city of Ames with about 80 different features in the data (two of the 82 features were IDs). I had worked with this dataset extensively withing MSDS 411, so I believe I am quite familiar with the data, and I would love to apply what we have learned in this class to that dataset and see what things we may be able to find.

With this data, I conducted an analysis of the features through unsupervised learning. For this dataset, I would first like to develop a way to best predict the house prices using the data from its features and find the features that would account for most of the variance of the house prices to be able to predict the prices relatively well. I would also like to see if similar price ranged houses have a similar feature set, whether that be location, number of rooms, lot size etc. What sort of segments would be identified within this data? What are their characteristics? Are there any anomalies? I would analyze trends in the data, and group together features that I would think are similar in relationship together or would follow similar trends. Then I would conduct feature analysis or EFA, to see if those features do end up grouped together. I would like to find anomalies within the dataset and use SOM to report those findings and what characteristics may have caused those certain data entries.

However, I would like to conduct an exploratory data analysis (or EDA) of this dataset beforehand to re-familiarize myself with the data and see if any changes need be made to further aid with the analysis.

## Methods and Findings of Exploratory Data Analysis (EDA):

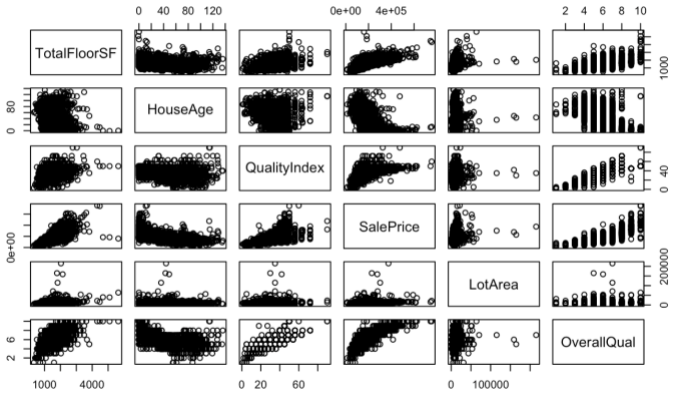
During my exploratory data analysis, my main objectives were to try and clean up the data as much as I could without overcomplicating the data and find some easy to spot trends that could help me later during my actual analysis.

The first step in my EDA was to create a few new variables to help with the analysis and may keep it simple. The following are the variables that I created:

- mydata$TotalFloorSF <- mydata$FirstFlrSF + mydata$SecondFlrSF
- mydata$HouseAge <- mydata$YrSold - mydata$YearBuilt
- mydata$QualityIndex <- mydata$OverallQual * mydata$OverallCond
- mydata$logSalePrice <- log(mydata$SalePrice)
- mydata$price_sqft <- mydata$SalePrice/mydata$TotalFloorSF

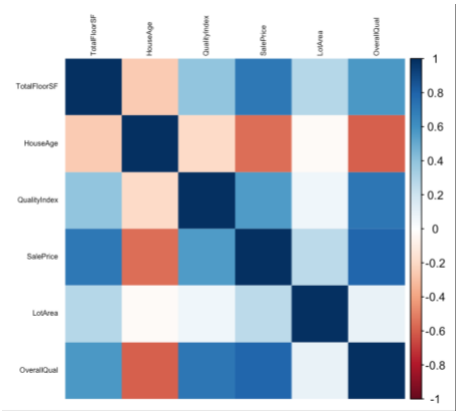
The log of the sale price and price per square foot are ones that I ended up not using much except for when doing some of the early EDA.

I then created some correlation plots to look for some easy to spot correlations.



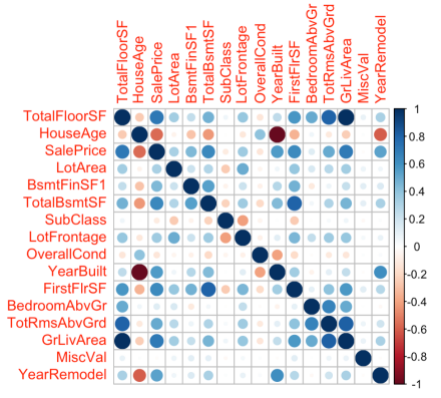

From this correlation plot the following observations were made:

- Overall Quality and Sale Price seems to have a bit of a linear relationship, as well as quality index but that is to be expected as higher quality homes tend to have a higher price.
- Sale Price and Total Floor Square Footage seems to have a slight linear relationship, indicating the price usually increases with the size of the house.
- Sale Price and House Age seem to have a slight opposite coordination.




From this correlation plot, the following new observations were made:

- Overall Quality and House age appear to have a strong inverse relationship. This is also affecting the house price.
- Lot area and house age are not in any relation, nor does it have any relation with the Quality Index.



From this correlation plot, the following inferences were made:

- The year built and Sale price are quite inversely related in this chart.
- First floor SF is quite in line with total basement square footage
- Year remodeled is inversely related to house age, which makes sense as older houses may need to be remodeled more as opposed to newer house.
- Many of the living area and floor square footage features are in line which each other as they go hand in hand quite a bit.

The first thing I did was look at how I could work with the data, especially because there were a lot of string features. To keep my analysis simple, I mainly focused on the numerical features of my data and did not really use the features that included string values, except for one exception which I thought was quite important. That exception was the "Neighborhood" feature, which I would work with by one hot encoding down the line. I removed most of the other string features as they were not so helpful with my analysis, would overcomplicate them, or could be taken into account within the "Neighborhood" feature. This "Neighborhood" feature would be converted into a factor to analyze against when conducting the SOM analysis to find anomalies.

## Data Preparation Description:

The first step in my EDA was to create a few new variables to help with the analysis and may keep it simple. The following are the variables that I created:

- mydata$TotalFloorSF <- mydata$FirstFlrSF + mydata$SecondFlrSF
- mydata$HouseAge <- mydata$YrSold - mydata$YearBuilt
- mydata$QualityIndex <- mydata$OverallQual * mydata$OverallCond

- mydata$logSalePrice <- log(mydata$SalePrice)
- mydata$price_sqft <- mydata$SalePrice/mydata$TotalFloorSF

This could help combine some elements of the data to make the models simpler. It could also help in determine if another feature may have a better relationship with a combined feature set rather than just be on its own.

After I had removed most of the string values, and the Overall Condition and Overall Quality values as they would be sort of repetitive to include and didn't seem to give me much insight, there were not a lot of rows of data entries left with missing values. Hence, if there was a missing value, I just removed it to avoid further complications within the analysis.

Due to the more simplistic nature of the data, I had decided to keep most of the data the same and keep all entries without focusing much on removing outliers, as anomalies are something I would like to investigate down the line in my analysis.

I did, however, end up standardizing the data down the line for my SOM analysis when conducting a search for the outliers so that the distributions of the data with respect to the factors I was finding anomalies against so that they can be compared on a similar grid and some trends could be more easily visible.

## Analysis Methods:

**Conducting Linear Regression:**

When conducting the linear analysis part of the experiment, I had done the data preparation discussed previously and created that newly created dataset "subdatanum," as it consisted of what I thought were the most important numerical features worth analyzing.

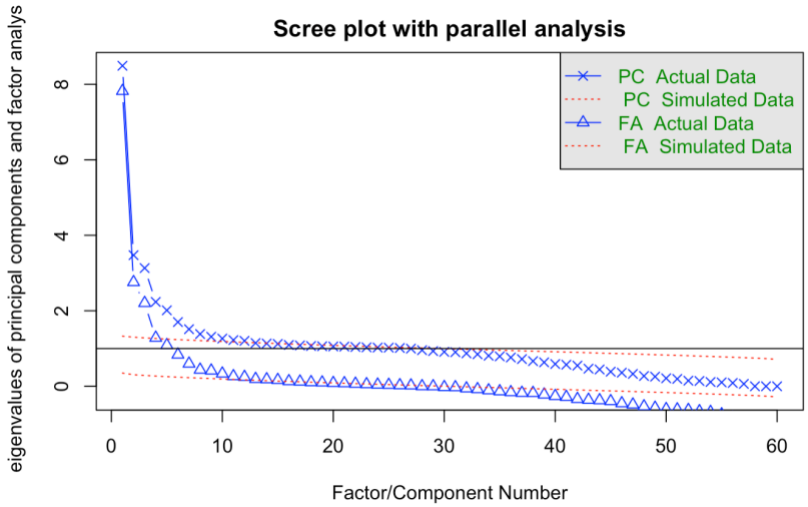
I had still maintained the original dataset as “mydata” to run a linear regression on just to see how that would turn out. Although this would be the most accurate, this had some “cheating variables” within it so was immediately discredited.

I also worked by picking out subsets of the data to see how much of the variance a lot of the features would carry with them for the data against the Sale Price feature. One of the best performing models was the model with TotalFloorSF, HouseAge, OverallQual, and LotArea (this was still in the early stages before I had removed OveralQual and OveralCond and replacing it with QualityIndex altogether). This is one example of how it would look like:

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + HouseAge + OverallQual +
    LotArea + MiscVal, data = subdata)

Residuals:
    Min      1Q  Median      3Q     Max
-446261  -21876   -2176   17569  282717

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.164e+04  4.625e+03 -13.326  < 2e-16 ***
TotalFloorSF  5.829e+01  1.822e+00  31.983  < 2e-16 ***
HouseAge     -4.803e+02  2.950e+01 -16.281  < 2e-16 ***
OverallQual   2.647e+04  7.519e+02  35.203  < 2e-16 ***
LotArea       1.164e+00  9.458e-02  12.303  < 2e-16 ***
MiscVal      -6.910e+00  1.260e+00  -5.485 4.49e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Feature Analysis (EFA):**

For the feature analysis segment of this experiment, I had used the best performing linear model I could create and analyze which features would be grouped together to account for most of the variance all while reducing the dimensionality of the data. I used a scree plot to determine how many factors I should retain of the data:




According to the scree plot, parallel analysis suggests that the number of factors = 15 and the number of components = 12.

After I conducted EFA with 15 factors, I then found which characteristics are grouped together, and roughly estimate whether the performance would be well by using the newly generated features in a linear regression via the root mean squared error value.

**SOM analysis to find anomalies:**

I will be conducting two SOM analyses within this experiment. The first one will be with respect to the "Neighborhood" feature, and I determined which neighborhood(s) contribute the most to the anomalies within the data. I created a distribution map of other features against the "Neighborhood" factor and see how the distributions lie and why such anomalies may have been caused.

I will train the SOM model for a certain grid size and for a certain amount of epochs to generate a counts plot and Neighborhood distance plot, which I will find the anomalies by finding which dots appear distant from the others (or have a very different color from those around it) and see

what entries from which neighborhood will make up that cluster. The same will be done with respect to Quality Index as the analyzing factor.

## Final Model Presentation:

**Conducting Linear Regression:**

For the best linear regression model, I used the following variables/features:

| | | | |
|---|---|---|---|
| "SubClass" | "FirstFlrSF" | "Fireplaces" | "HouseAge" |
| "LotFrontage" | "SecondFlrSF" | "GarageYrBlt" | "Neighborhood" |
| "LotArea" | "LowQualFinSF" | "GarageCars" | |
| "OverallCond" | "BsmtFullBath" | "GarageArea" | |
| "YearBuilt" | "BsmtHalfBath" | "WoodDeckSF" | |
| "YearRemodel" | "FullBath" | "OpenPorchSF" | |
| "MasVnrArea" | "HalfBath" | "ScreenPorch" | |
| "BsmtFinSF1" | "BedroomAbvGr" | "PoolArea" | |
| "BsmtFinSF2" | "KitchenAbvGr" | "MiscVal" | |
| "BsmtUnfSF" | "TotRmsAbvGrd" | "TotalFloorSF" | |

All of these features were analyzed to predict "SalePrice."

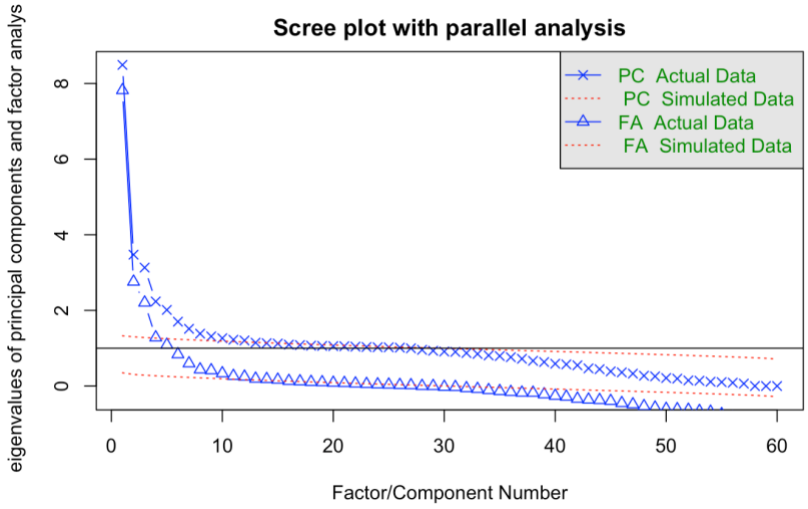
The following are the results of the model:

```
Residual standard error: 32260 on 2216 degrees of freedom
Multiple R-squared:  0.8539,    Adjusted R-squared:  0.8501
F-statistic: 227.2 on 57 and 2216 DF,  p-value: < 2.2e-16
```

I had gotten a relatively high R^2 value of 0.8501, which is quite high for not including a lot of the original variables, and especially not the “cheating variables” that had to do with price. These features will be used for the upcoming feature analysis segment of the experiment.

**Feature Analysis (EFA):**

I used a scree plot to determine how many factors I should retain of the data:




According to the scree plot, parallel analysis suggests that the number of factors = 15 and the number of components = 12.

After I conducted EFA with 15 factors, I then found which characteristics are grouped together, and roughly estimate whether the performance would be well by using the newly generated features in a linear regression via the root mean squared error value. I used a cutoff value of 0.5 to determine the features each factor should retain. These are the features in each factor:

- Factor 1: YearBuilt + YearRemodel + GarageYrBuilt + HouseAge
- Factor 2: GarageCars + GarageArea
- Factor 3: SecondFlrSF + HalfBath
- Factor 4: BsmtFinSF1 + BsmtUnfSF + BsmtFullBath
- Factor 5: BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd
- Factor 6: SubClass + LotFrontage

- Factor 8: Fireplaces
- Factor 9: Neighborhood_NridgHt
- Factor 10: Neighborhood_Names
- Factor 11: Neighborhood_Somerst
- Factor 12: Neighborhood_OldTown
- Factor 13: MasVnrArea
- Factor 15:  OverallCond

And here are the models with the values corresponding to the factors with cutoff >0.5:

- M 1: 0.872(YearBuilt) + 0.510(YearRemodel) + 0.664(GarageYrBlt) - 0.879(HouseAge)
- M 2: 0.861(GarageCars) + 0.908(GarageArea)
- M 3: 0.828(SecondFlrSF) + 0.650(HalfBath)
- M 4: 0.735(BsmtFinSF1) - 0.792(BsmtUnfSF) + 0.714(BsmtFullBath)
- M 5: 0.704(BedroomAbvGr) + 0.518(KitchenAbvGr) + 0.642(TotRmsAbvGrd)
- M 6: 0.834(SubClass) - 0.583(LotFrontage)
- M 8: 0.563(Fireplaces)
- M 9: 1.036 (Neighborhood_NridgHt)
- M 10: -1.003(Neighborhood_Names)
- M 11: 0.999(Neighborhood_Somerst)
- M 12: 1.007(Neighborhood_OldTown)
- M 13: 0.586(MasVnrArea)
- M 15: 0.542(OverallCond)

Some of the factors are grouped together quite well, as factor 1 has all time-based variables, factor 2 has garage-based variables, factor 4 has basement-based values, and factor 5 has above ground variables. The rest are quite interesting, as many of what I would consider different variables are grouped together, such as Sub Class and Lot Frontage, as well as Second Floor SF and Half Bathrooms. The rest of the features weren't so alike and hence were separated.

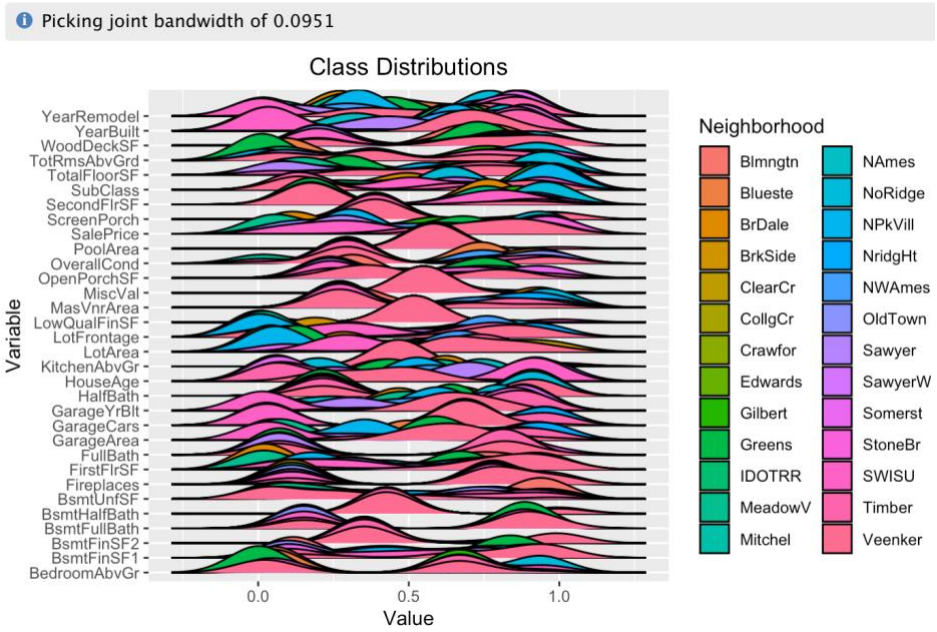
Below are the statistics of the factor analysis:

```
Mean item complexity =  3.2
Test of the hypothesis that 15 factors are sufficient.

The degrees of freedom for the null model are  1770  and the objective function was  99.07
The degrees of freedom for the model are 975  and the objective function was  71.27

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  0.03

Fit based upon off diagonal values = 0.98
```

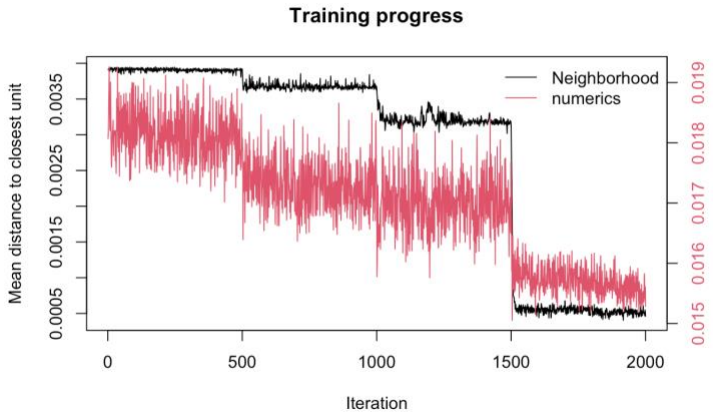
The model with 15 factors tends to be the best performing model. The root mean squared value is very low at 0.02. It appears to account for the most variance for this model all while reducing the dimensionality of the data by 30 variables. Even in this model, we would only use 13 of the factors as models due to two of them being insignificant (factor 7 and 14, as no values are above 0.5 for any of the features). The scree analysis for this experiment was almost perfect.

SOM Analysis:

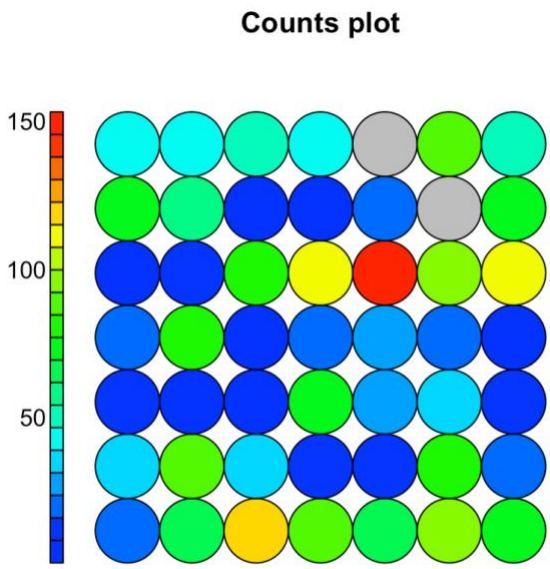
The first SOM analysis was run with respect to Neighborhood. Here is the distribution map of features with respect to neighborhood:



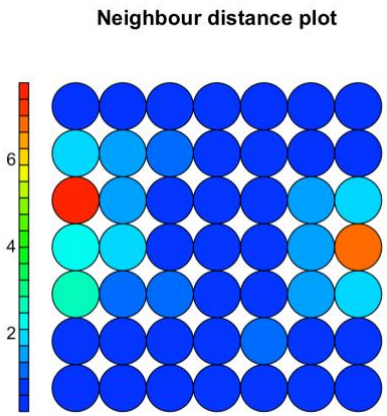
Afterwards I ran the SOM model with a similar output that we had for our assignment in class, which was for 2,000 epochs, but I modified its grid size to a 7 by 7 grid, as the original 15 by 15 grid returned too many empty grey circles within the Counts Plot.




The 7 by 7 grid resulted in only 2 grey dots on the counts plot:




To find the clusters that contain anomalies, we look for clusters with extremely different colors than from those around it in the Neighborhood distance plot:




From the neighbor distance plot, it appears the red dot on the left and the orange dot on the right are the most distant from those around it. That is where we would be able to find our anomalies:

| WoodDeckSF <dbl> | OpenPorchSF <dbl> | ScreenPorch <dbl> | PoolArea <dbl> | MiscVal <dbl> | SalePrice <dbl> | TotalFloorSF <dbl> | HouseAge <dbl> | Neighborhood <fctr> |
|---|---|---|---|---|---|---|---|---|
| 0.10929671 | -0.715245909 | -0.2949381 | -0.06874654 | -0.08405242 | -0.8925764 | -1.36727827 | 0.3641893 | IDOTRR |
| -0.75681270 | 0.740709658 | -0.2949381 | -0.06874654 | -0.08405242 | -0.7065875 | -0.69099055 | 1.6953854 | IDOTRR |
| -0.75681270 | -0.715245909 | -0.2949381 | -0.06874654 | -0.08405242 | -1.2285563 | -1.91951321 | 1.0784896 | IDOTRR |
| -0.75681270 | 0.134061505 | 3.1612431 | -0.06874654 | -0.08405242 | -0.8925764 | -0.78503055 | 1.7278536 | IDOTRR |
| -0.51168739 | -0.715245909 | -0.2949381 | -0.06874654 | -0.08405242 | -1.3917465 | -0.74301268 | 1.7603218 | IDOTRR |
| -0.20119534 | -0.715245909 | -0.2949381 | -0.06874654 | -0.08405242 | -0.9792592 | -0.64297012 | 2.4096858 | IDOTRR |
| -0.75681270 | -0.715245909 | -0.2949381 | -0.06874654 | 0.02167803 | -1.5406216 | -0.91108418 | 0.7213394 | IDOTRR |
| -0.75681270 | -0.715245909 | -0.2949381 | -0.06874654 | -0.08405242 | -1.6108774 | -1.09516250 | 2.5395586 | IDOTRR |
| -0.75681270 | -0.715245909 | -0.2949381 | -0.06874654 | -0.08405242 | -0.4246044 | 0.86167002 | 1.5005762 | IDOTRR |
| -0.75681270 | -0.715245909 | -0.2949381 | -0.06874654 | -0.08405242 | -0.5925943 | 0.33344529 | 1.5005762 | IDOTRR |

1–10 of 64 rows | 26–34 of 33 columns

Previous 1 2 3 4 5 6 7 Next

| WoodDeckSF <dbl> | OpenPorchSF <dbl> | ScreenPorch <dbl> | PoolArea <dbl> | MiscVal <dbl> | SalePrice <dbl> | TotalFloorSF <dbl> | HouseAge <dbl> | Neighborhood <fctr> |
|---|---|---|---|---|---|---|---|---|
| -0.7568127 | -0.7152459 | -0.2949381 | -0.06874654 | -0.08405242 | 2.287233 | 1.07376 | 0.2667847 | Timber |

1 row | 26–34 of 33 columns

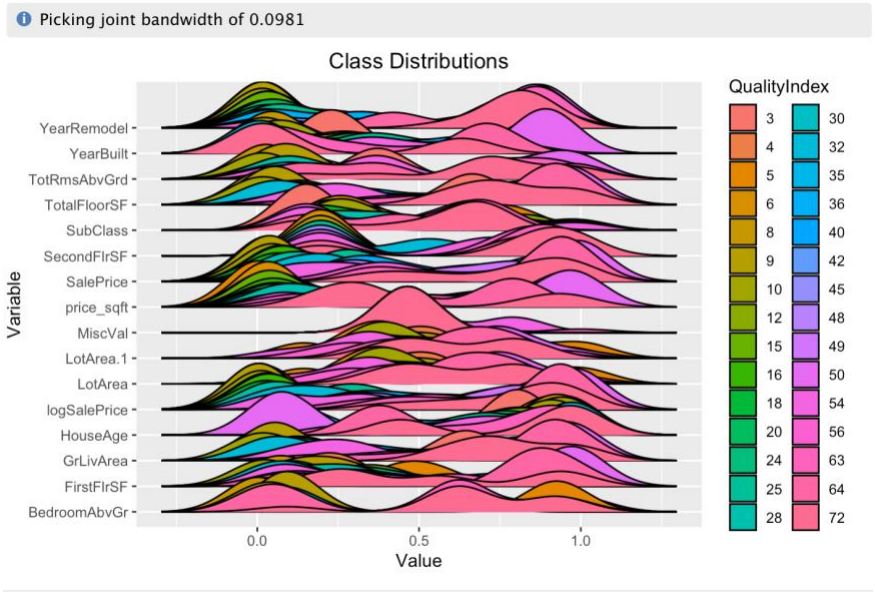
There are a total of 65 anomalies within this dataset. The first 64 rows consist of those from the red dot on the Neighbor distance plot and the 65th one comes from the orange dot on the same plot.

The neighborhoods these anomalies come from are IDOTRR, consisting of 64 of the 65 anomalies, with the last one being from the neighborhood Timber. It could be that IDOTRR has a very diverse subset of properties, consisting of houses with a diverse feature set, The neighborhood comes as green on the distribution map when being analyzed, which has a very wide distribution or is usually split up onto both ends of the map.  For example, there are many

houses with a few bedrooms above ground and many with a lot. It could have a newly developing area as well as an old existing one.

Timber has one house that just has an extremely higher sale price than the usual houses in the area. This could be a custom-built house or a property classified as a mansion.

Next, I will conduct a SOM analysis against the “QualityIndex” feature, and I determined which index values contributed most to the anomalies within the data and compare that against the distribution map generated from the features with respect to the Quality Index to determine what might make up the anomalies.

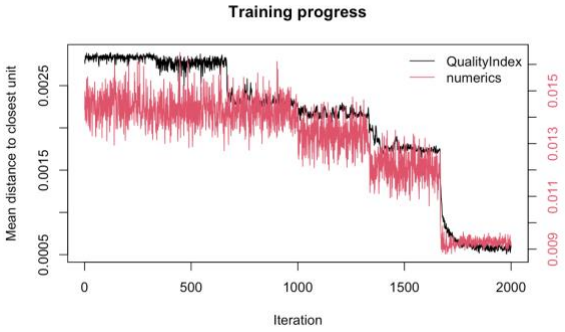



Some trends I found right off the bat:

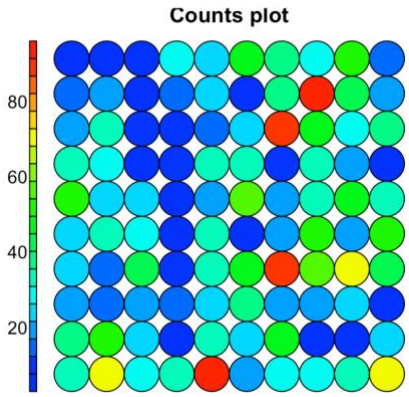
- Low quality houses on the lower end of the year remodeled distribution. This could be as a lot of houses may not have been remodeled and could be deteriorating.
- In the second-floor square footage, lower quality houses seem to be on the lower end. This may be because of the abundance of smaller houses who may not have any second floor. Smaller houses are often cheaper and could be of lower quality.

- Lower quality houses spike with a greater number of bedrooms above ground. Maybe this results from other housing units that are more commercial and tend to be more heavily abused than private houses.
- There is a large number of older houses with greater quality index values, which is surprising as they are usually the ones that are deteriorating nowadays and should be on the opposite end.

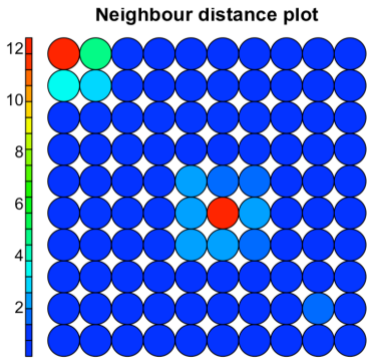
Afterwards I ran the SOM model with a similar output that we had for our assignment in class, which was for 2,000 epochs, but I modified its grid size to a 10 by 10 grid, as the original 15 by 15 grid size returned too many empty grey circles within the Counts Plot.




The 10 by 10 grid resulted in no grey dots on the counts plot:



To find the clusters that contain anomalies, we look for clusters with extremely different colors than from those around it in the neighborhood distance plot:




The two red dots here are of concern. That is where we would find the anomalies within our data. Overall, the data seems to be quite consistent on this end.

| | TotalFloorSF <dbl> | HouseAge <dbl> | QualityIndex <fctr> | price_sqft <dbl> | SalePrice <dbl> | LotArea <dbl> | logSalePrice <dbl> | LotArea.1 <dbl> |
|---|---|---|---|---|---|---|---|---|
| 39 | 0.88443584 | -1.1697769 | 45 | 2.5745833 | 2.6837504 | 0.001405854 | 2.1250886 | 0.001405854 |
| 42 | 0.40344699 | -1.0377260 | 45 | 1.2654704 | 1.1792194 | 0.174121201 | 1.2354609 | 0.174121201 |
| 45 | 1.72716011 | -1.1697769 | 45 | 4.3000518 | 5.3934007 | 0.351658872 | 3.1967617 | 0.351658872 |
| 49 | 0.51077508 | -1.0377260 | 45 | 1.9125048 | 1.7412655 | -0.315979217 | 1.6065163 | -0.315979217 |
| 322 | 1.03946531 | -1.1367642 | 45 | 2.0683561 | 2.4748044 | 0.389095336 | 2.0192078 | 0.389095336 |
| 348 | 0.49884973 | -1.1037515 | 45 | 2.9665543 | 2.4622867 | 0.602039018 | 2.0127171 | 0.602039018 |
| 367 | 1.46877767 | -1.1697769 | 45 | 3.2307977 | 4.0187036 | 0.923231188 | 2.7112302 | 0.923231188 |
| 425 | 0.31201935 | -1.1037515 | 45 | 2.3558548 | 1.8051059 | 0.383892302 | 1.6453221 | 0.383892302 |
| 430 | 1.04344042 | -1.1697769 | 45 | 2.4583454 | 2.7940065 | 0.279450913 | 2.1791707 | 0.279450913 |
| 435 | 0.47897416 | -1.1697769 | 45 | 2.6895307 | 2.2432264 | 0.181481591 | 1.8962596 | 0.181481591 |

1–10 of 69 rows | 1–9 of 17 columns

Previous 1 2 3 4 5 6 7 Next

| | TotalFloorSF <dbl> | HouseAge <dbl> | QualityIndex <fctr> | price_sqft <dbl> | SalePrice <dbl> | LotArea <dbl> | logSalePrice <dbl> | LotArea.1 <dbl> | SubClass <dbl> |
|---|---|---|---|---|---|---|---|---|---|
| 254 | 3.432484 | -0.5425353 | 35 | -0.6985153 | 1.7425173 | 0.2827504 | 1.607283 | 0.2827504 | 0.0612746 |
| 566 | 3.001184 | -1.0377260 | 35 | -0.8831800 | 1.2511964 | -0.9223230 | 1.286232 | -0.9223230 | 2.4065990 |
| 816 | 2.567897 | -0.9056752 | 35 | -0.7806087 | 1.1103719 | 0.2166338 | 1.185894 | 0.2166338 | 0.7648719 |
| 817 | 2.567897 | -0.9056752 | 35 | -0.7806087 | 1.1103719 | -0.2803194 | 1.185894 | -0.2803194 | 0.7648719 |
| 818 | 2.567897 | -0.9056752 | 35 | -0.7806087 | 1.1103719 | -0.2756240 | 1.185894 | -0.2756240 | 0.7648719 |
| 2273 | 1.969642 | -1.0707387 | 35 | -0.2823370 | 1.2405563 | 0.2096541 | 1.278792 | 0.2096541 | 0.0612746 |
| 2351 | 2.561934 | -0.2784335 | 35 | -0.9969583 | 0.8662762 | 0.8304649 | 1.001621 | 0.8304649 | 0.0612746 |

7 rows | 1–10 of 17 columns

As you can see, there are a total of 86 total anomalies, and they appear to be in the Quality index value of 45 and 35, which are classified in the mid-range of the quality index. These two could have attributes contributing to being anomalies such that some features of the property may be in

great condition while others may be on the opposite end, or in bad or terrible condition. For example, the interior could be great, while the exterior may be terrible or vice versa, or it may be in a commercial in a bad location with bad exterior but great interior. These could be extreme cases on either end, hence making them anomalies within this data with respect to quality.

## Conclusion and Reflection:

Overall, this experiment was an enjoyable one. It was great having the freedom to finally create and do an entire experiment on your own using what you have learned in class. Still, if I had time, there are some more things I would like to have done.

I would have liked to conduct a Linear Estimation of the sale price with respect to the newly created factors. I would have like to see how well the data would have performed in predicting the sale price with the newly generated features. Would the performance had remained the same or changed with that?

I would have also liked to look in depth as to why these anomalies were anomalies. There could have been more specific reasons why those neighborhoods and quality index values had created so many anomalies. The analysis I did was a general idea of what I though the anomalies can be based on my thought. I would have loved to go into the specifics, and perhaps looked into my findings with someone familiar with the area and could give me a run down as to why these anomalies were found.