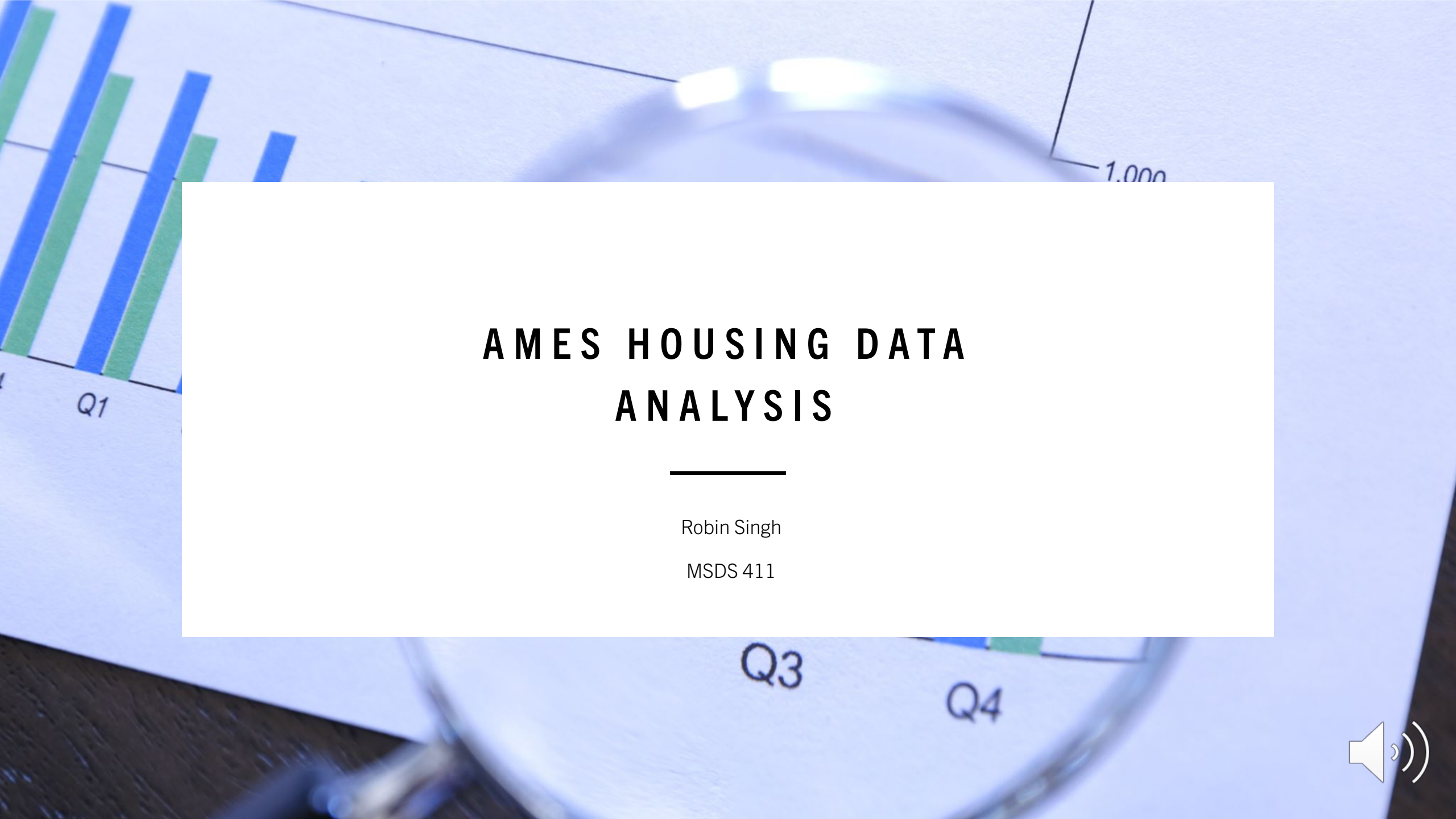


AMES HOUSING DATA ANALYSIS

Robin Singh

MSDS 411

ABOUT THE DATA

- For the experiment, I will be using the Ames Housing Dataset. This dataset contains 2930 records for real home properties listed in the housing market for that city. It lists various house features such as price, lot size, bedrooms, square footage, street, area, etc.

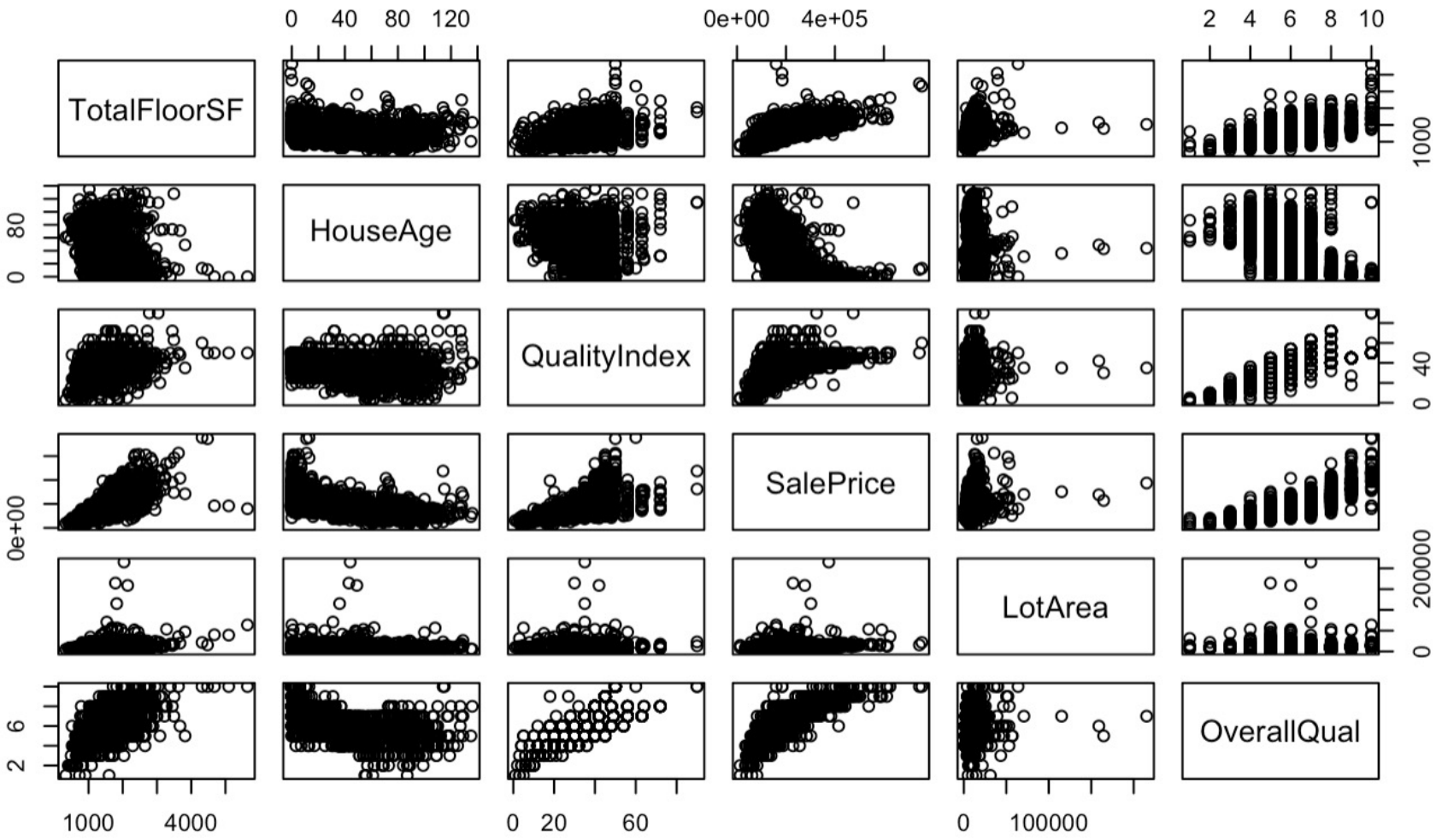
INTRODUCTION

- Begin with best linear model to predict the housing prices.
  - Find which features generate the most impact when predicting house prices.
  - What houses are grouped together (location, room number, lot size, etc.)
- Analyze the features and which features are grouped together via feature analysis or EFA.
- SOM will be used to find anomalies within the data.

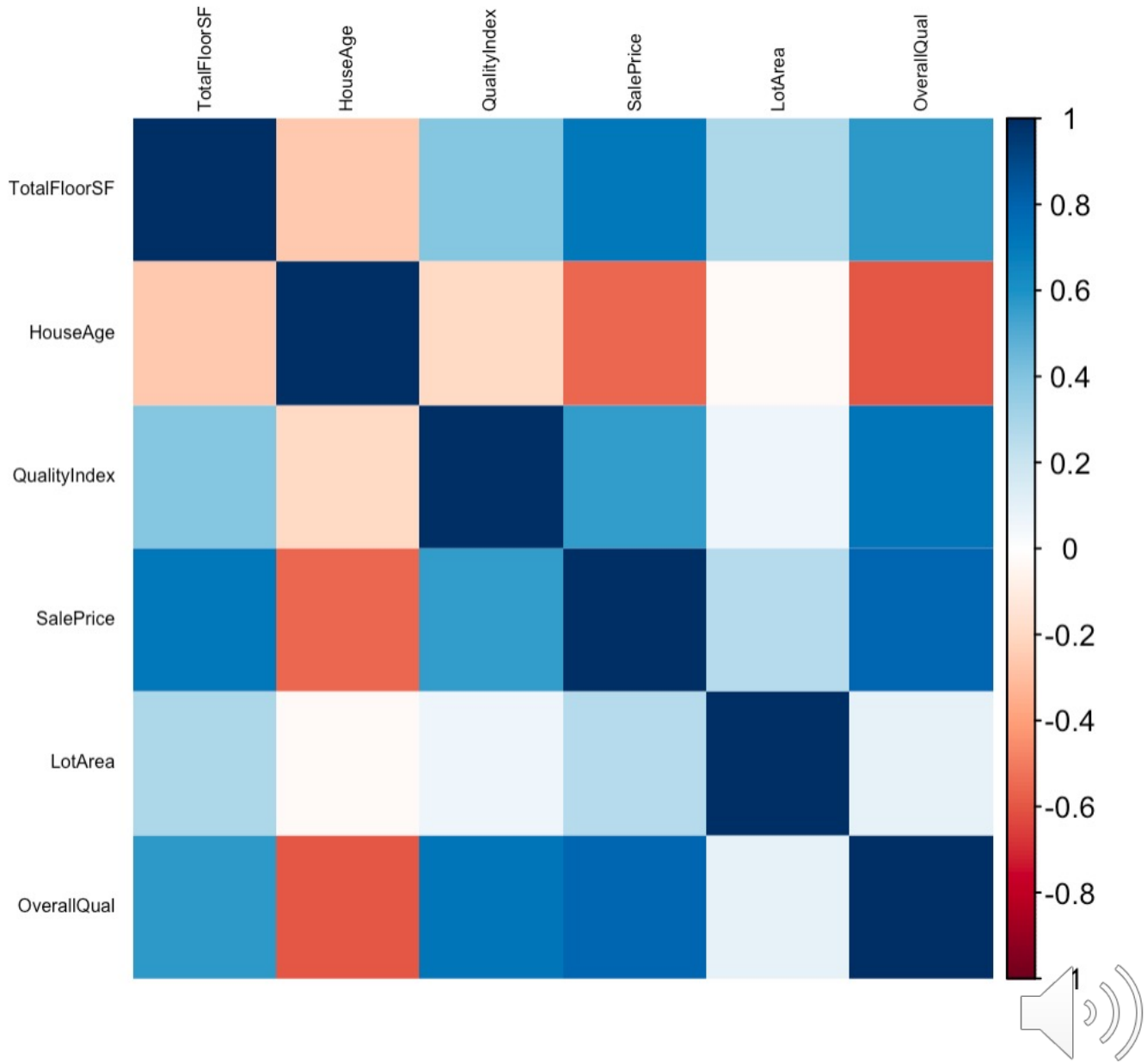
CREATED VARIABLES (EDA)

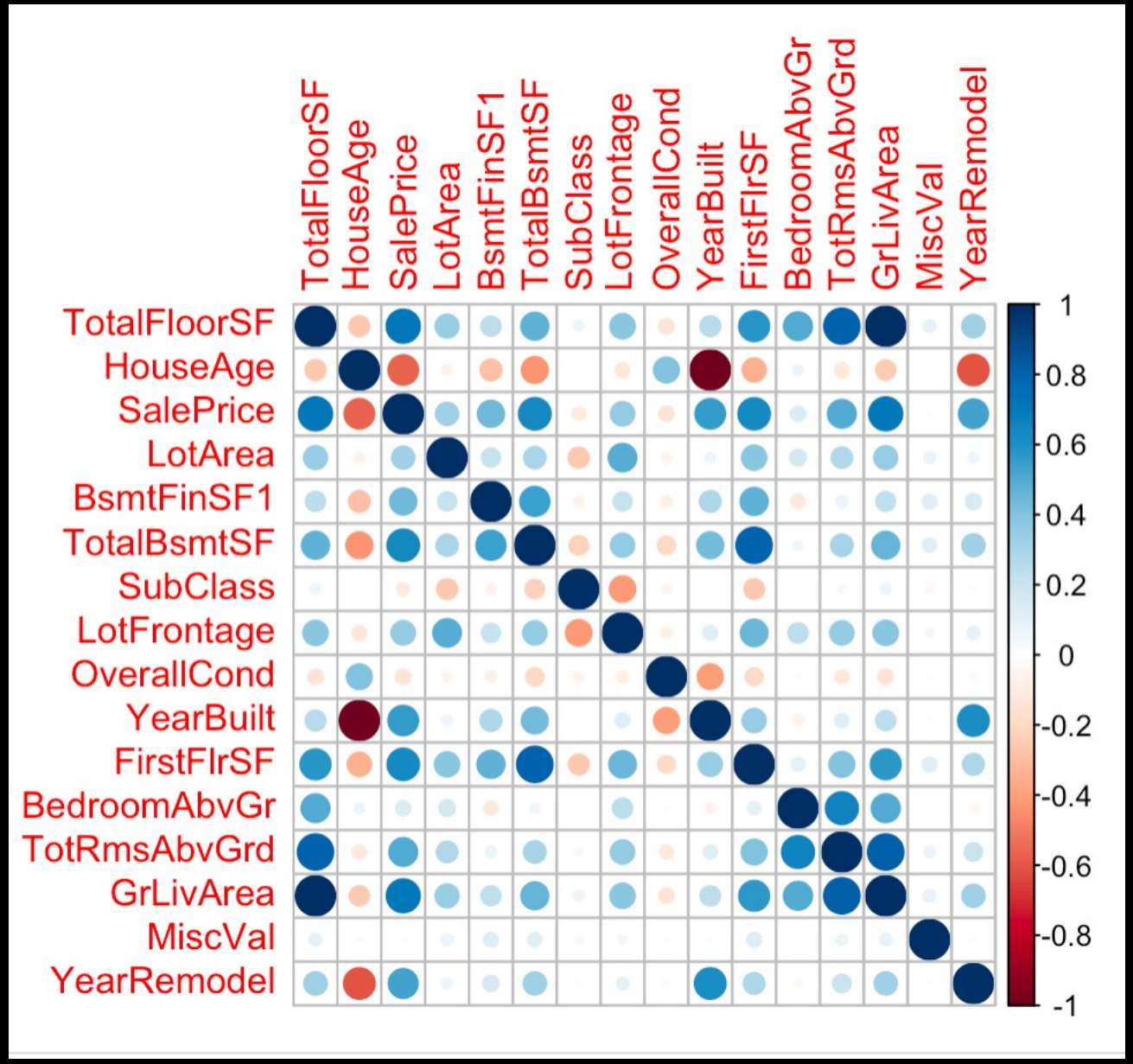
- mydata$TotalFloorSF <- mydata$FirstFlrSF + mydata$SecondFlrSF
- mydata$HouseAge <- mydata$YrSold - mydata$YearBuilt
- mydata$QualityIndex <- mydata$OverallQual * mydata$OverallCond
- mydata$logSalePrice <- log(mydata$SalePrice)
- mydata$price_sqft <- mydata$SalePrice/mydata$TotalFloorSF

EDA


TotalFloorSF
HouseAge
QualityIndex
SalePrice
LotArea
OverallQual
0 40 80 120
0e+00 4e+05
2 4 6 8 10
0 80
0 40
0e+00
0 200000
2 6
1000 4000
0 20 60
0 100000
1000

EDA (CONTINUED)




TotalFloorSF
HouseAge
SalePrice
LotArea
BsmtFinSF1
TotalBsmtSF
SubClass
LotFrontage
OverallCond
YearBuilt
FirstFlrSF
BedroomAbvGr
TotRmsAbvGrd
GrLivArea
MiscVal
YearRemodel
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

SUMMARY OF LINEAR MODEL 1

- Model 1: All variables included.
- Although most successful model with the highest R-Squared value retained, there were some issues.
- House Styles all returned N/A values, and hence should not be included in the model, as it does not seem to contribute to the accuracy levels.
- The lot shape seemed to have no impact on the accuracy, hence was unnecessary.
- Lot Area and second floor square footage was also unnecessary, and not being properly analyzed in the linear model.

SUMMARY OF LINEAR MODEL 2

- This was a much simpler model, which was run with Overall Quality, Total Square Footage, House Age, Miscellaneous Value, and Lot Area.
- These variables alone accounted for 76.5% of the variance in the linear model, making these some of the most important factors worth analyzing.

BEST LINEAR PREDICTION 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
```

- The best model included most numerical features, but also hot encoded neighborhood feature (since it was a string).
- List of features:

"SubClass" "LotFrontage" "LotArea" "OverallCond" "YearBuilt" "YearRemodel"

"MasVnrArea" "BsmtFinSF1" "BsmtFinSF2" "BsmtUnfSF" "FirstFlrSF" "SecondFlrSF"

"LowQualFinSF" "BsmtFullBath"

"BsmtHalfBath" "FullBath" "HalfBath" "BedroomAbvGr" "KitchenAbvGr" "TotRmsAbvGrd"

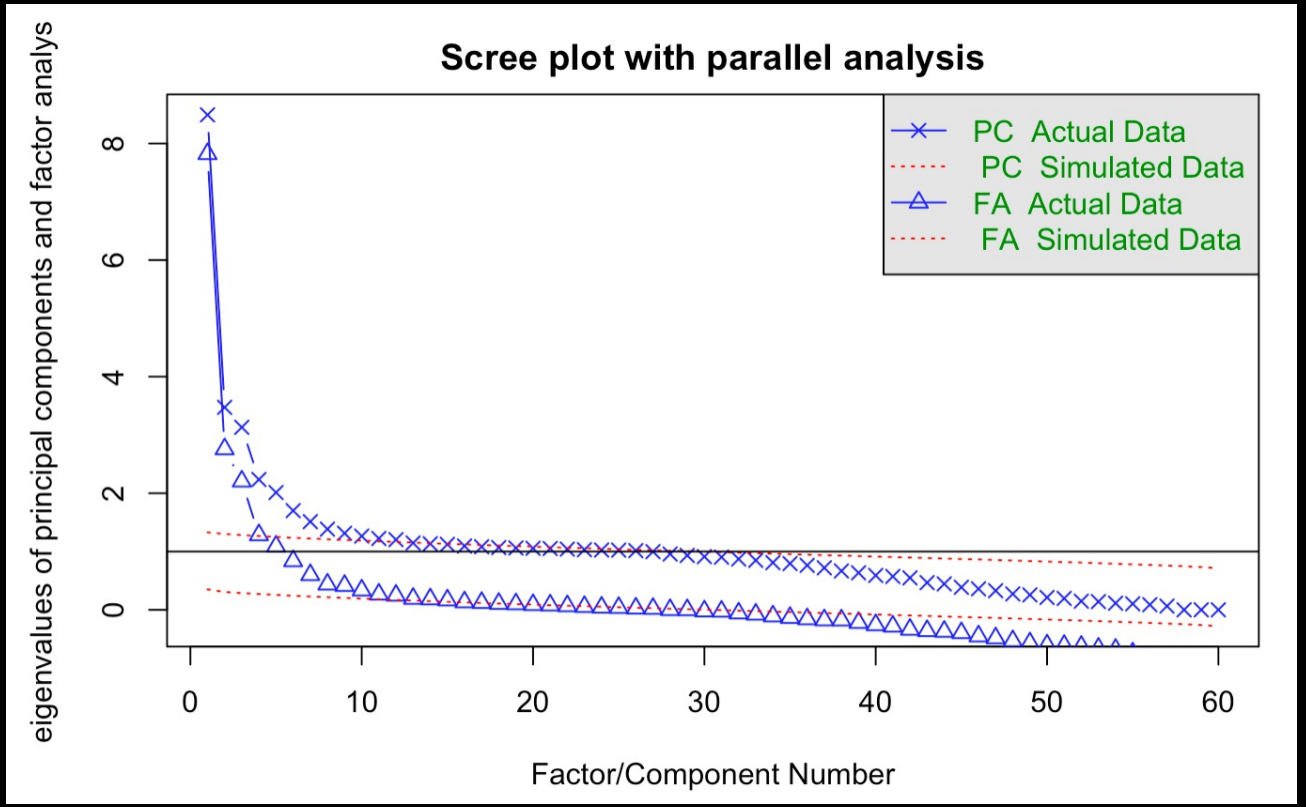
"Fireplaces" "GarageYrBlt"

"GarageCars" "GarageArea" "WoodDeckSF" "OpenPorchSF" "ScreenPorch" "PoolArea" "MiscVal"

"TotalFloorSF" "HouseAge" "Neighborhood"

Analyzed against "SalePrice"

CONDUCTING FEATURE ANALYSIS



Parallel analysis suggests that the number of factors = 15 and the number of components = 12

FACTOR ANALYSIS BREAKDOWN(>0.5)

- 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

MODELS

- 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)

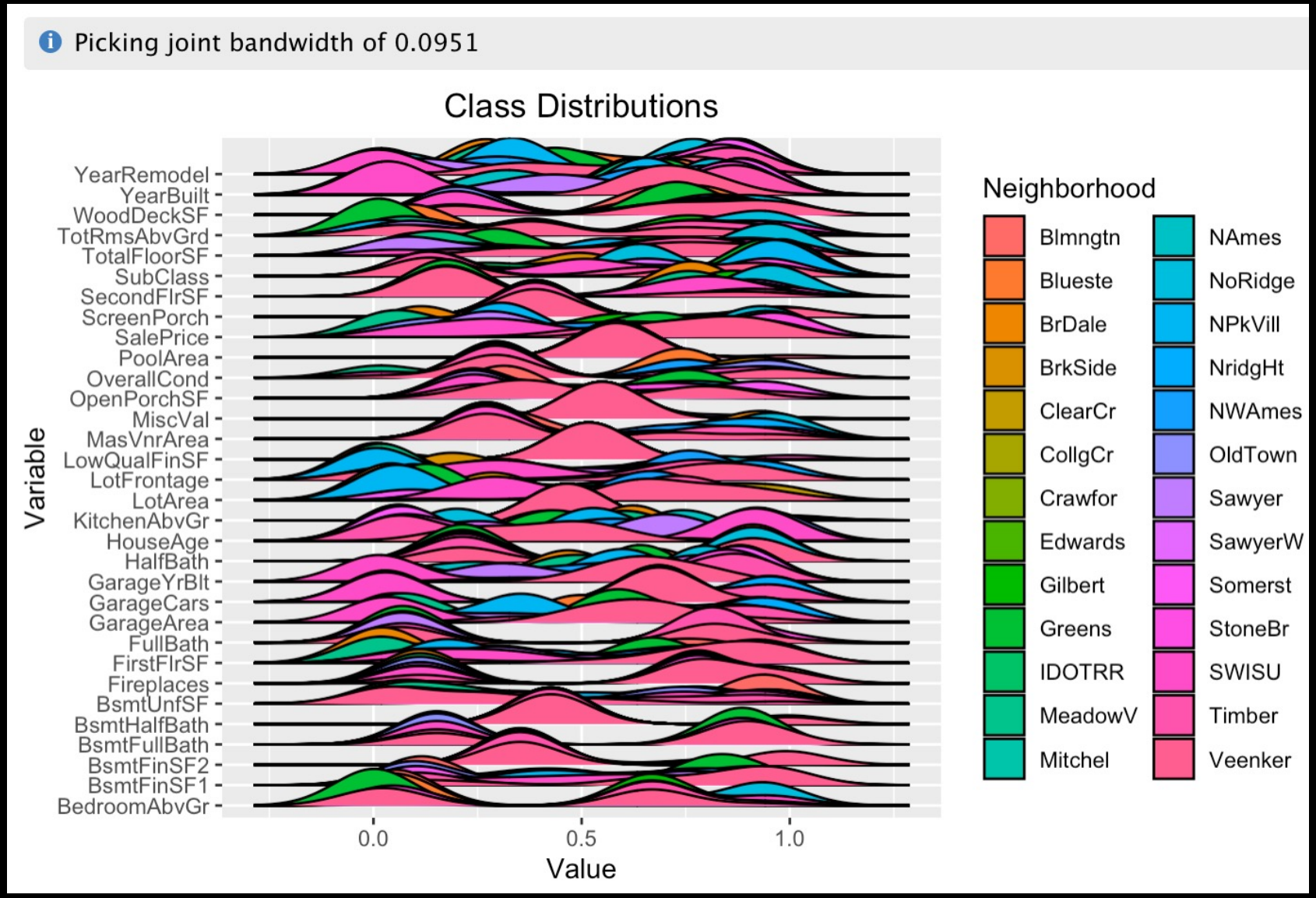
EXPLAINED

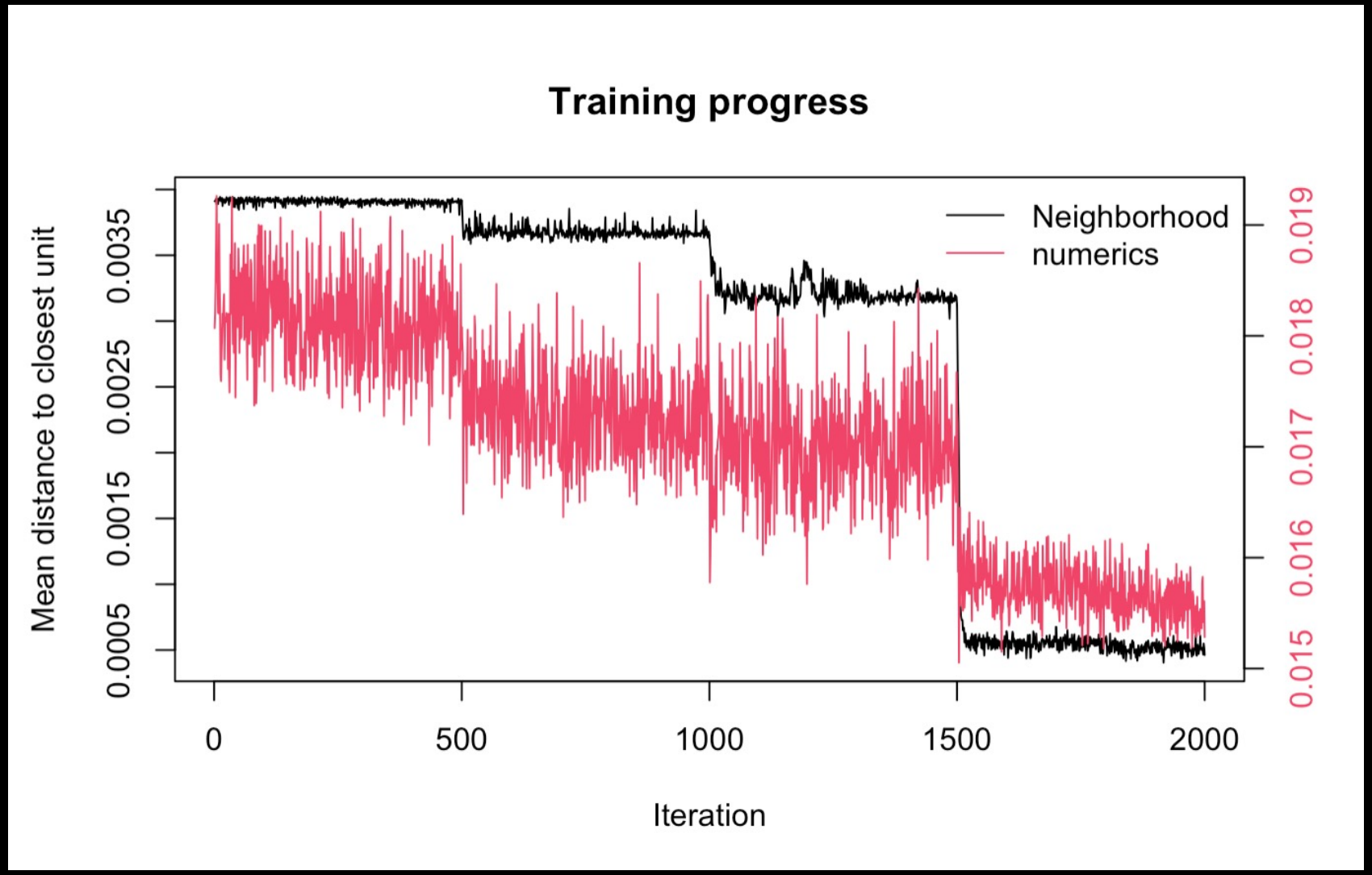
- The model with 15 factors tends to be the best performing model, as it has the highest designated p-value at the end, meaning it has the greatest chance of the null hypothesis being rejected. It appears to account for the most variance for this model all while reducing the dimensionality of the data 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 and14, as no values are above 0.5 for any of the features).
- The scree analysis for this experiment was almost

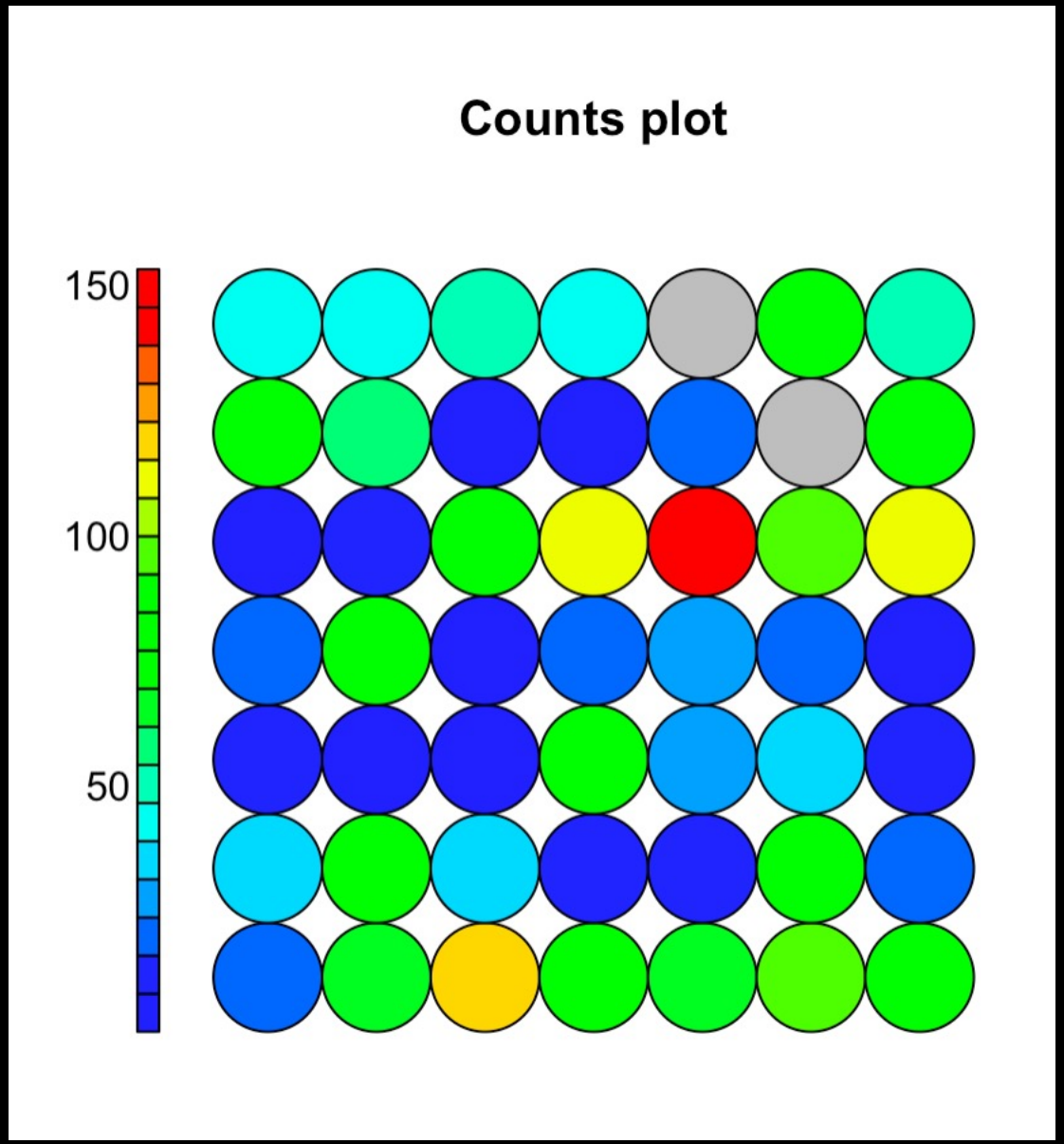
```
Mean item complexity =  3.2
Test of the hypothesis that 15 factors are sufficient.

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
```

CHECKING FOR ANOMALIES VIA SOM (NEIGHBORHOODS)

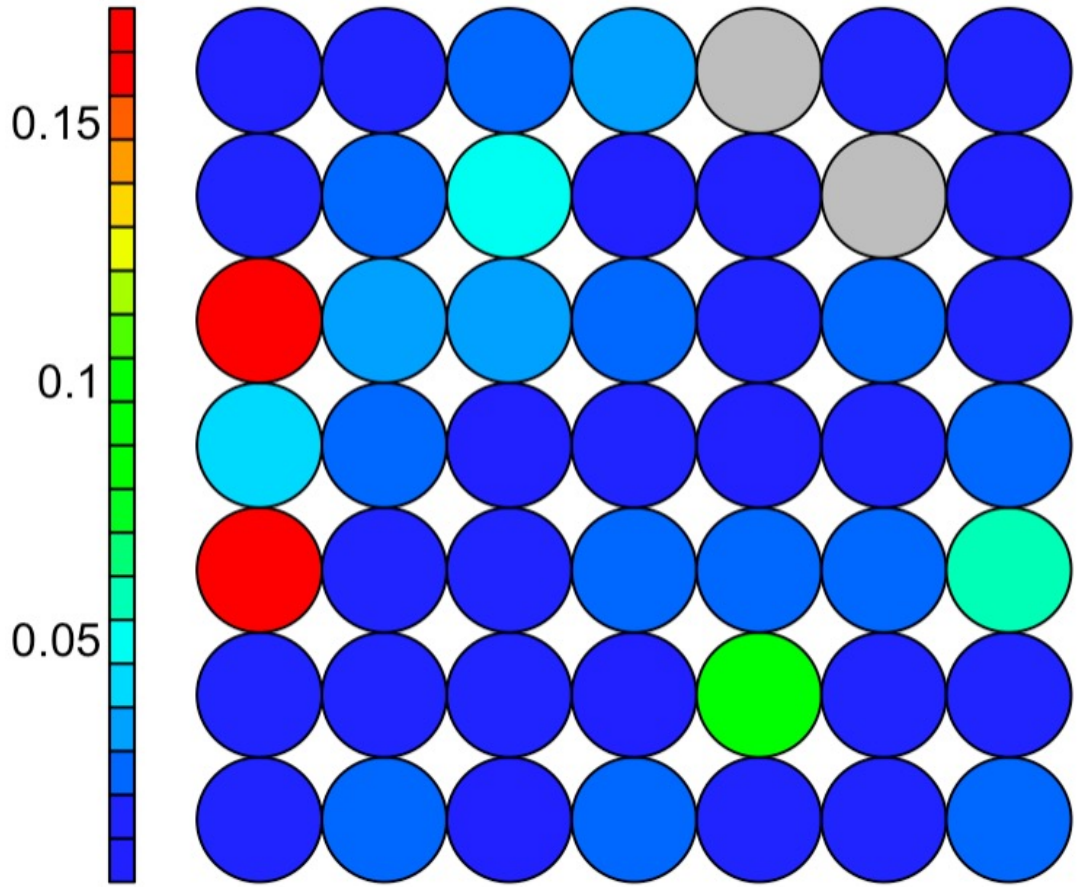
CHECK FOR ANOMALIES

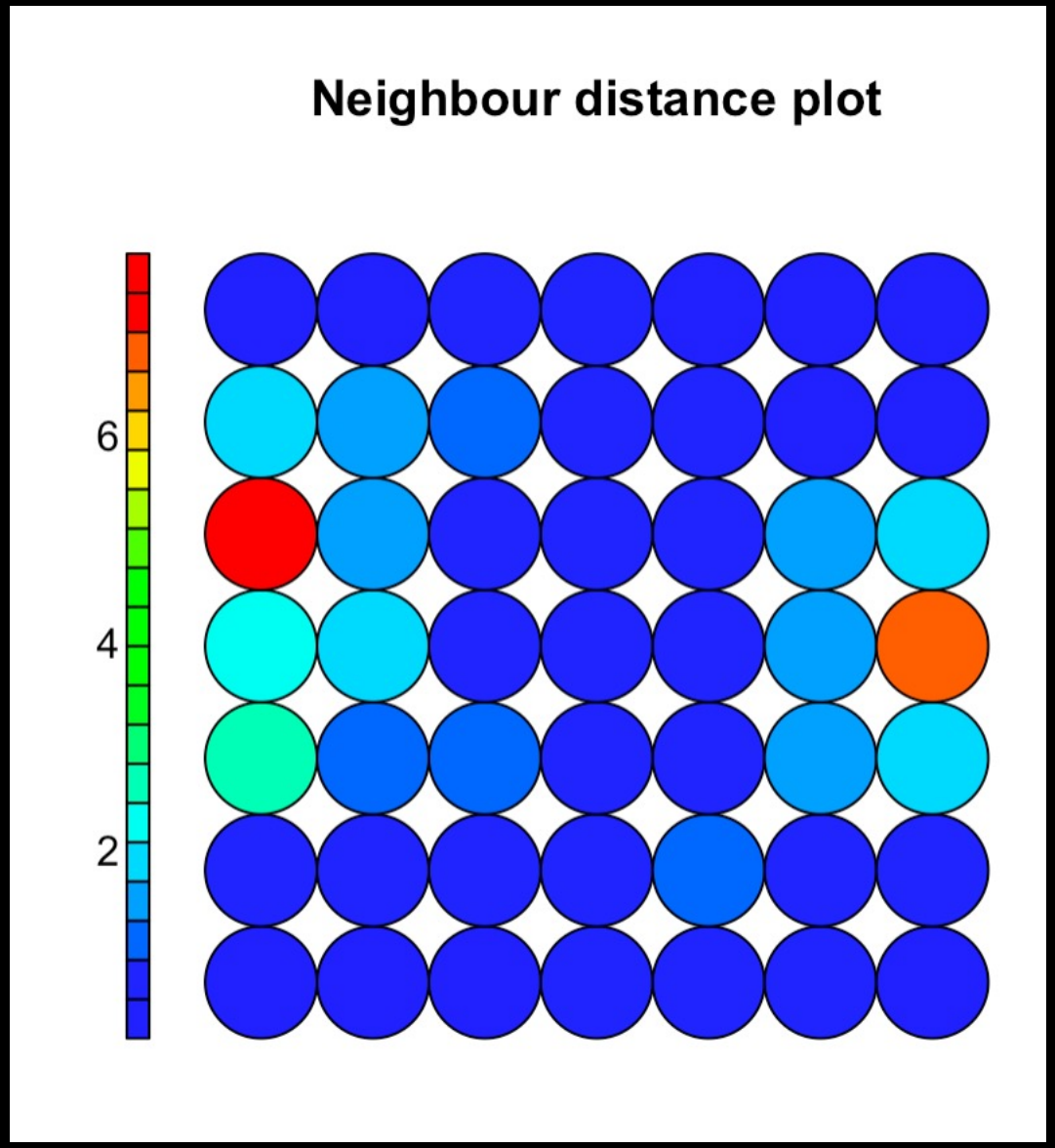



Training progress
Mean distance to closest unit
0.0005
0.0015
0.0025
0.0035
0.015
0.016
0.017
0.018
0.019
0
500
1000
1500
2000
Iteration
Neighborhood
numerics


Counts plot
150
100
50

Quality plot




Neighbour distance plot
6
4
2

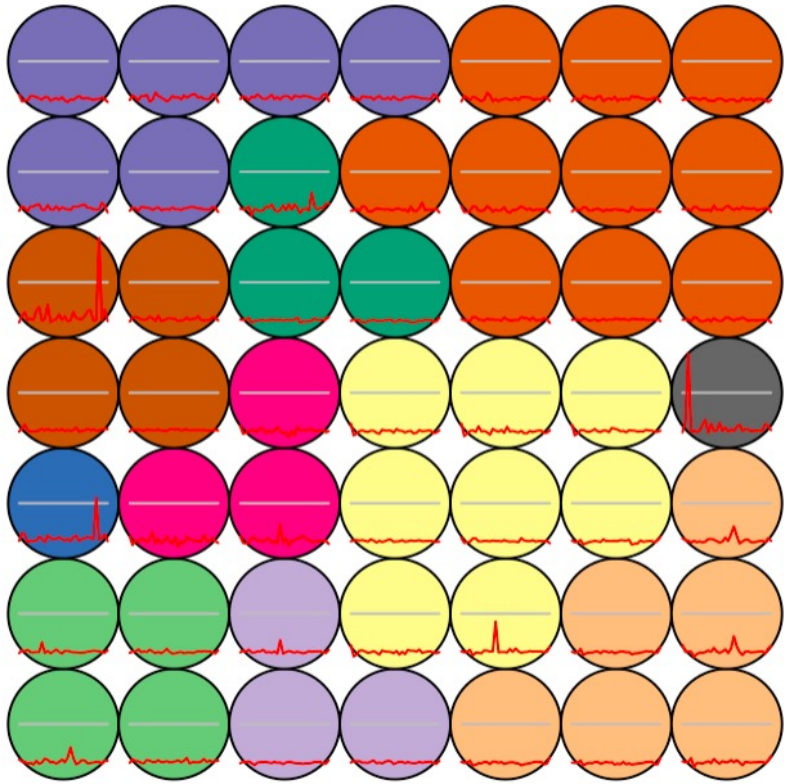
ANOMALIES

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

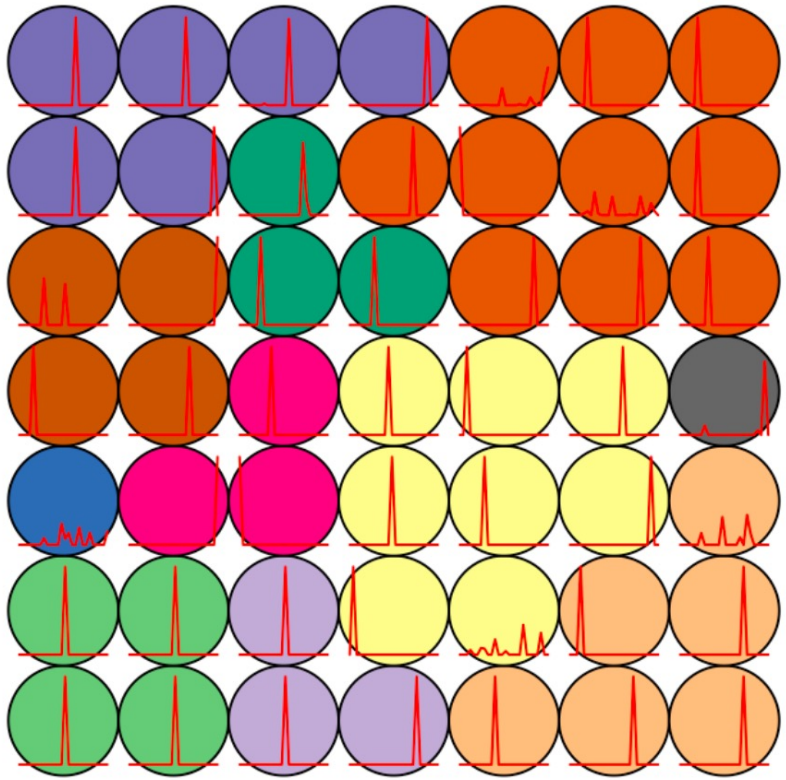
WoodDeckSF <dbl>	OpenPorchSF <dbl>	ScreenPorch <dbl>	PoolArea <dbl>	MiscVal <dbl>	SalePrice <dbl>	TotalFloorSF <dbl>	HouseAge <dbl>	Neighborhood <fctr>
-0.7568127	-0.7152459	-0.2949381	-0.06874654	-0.08405242	2.287233	1.07376	0.2667847	Timber

1 row | 26-34 of 33 columns

Clusters

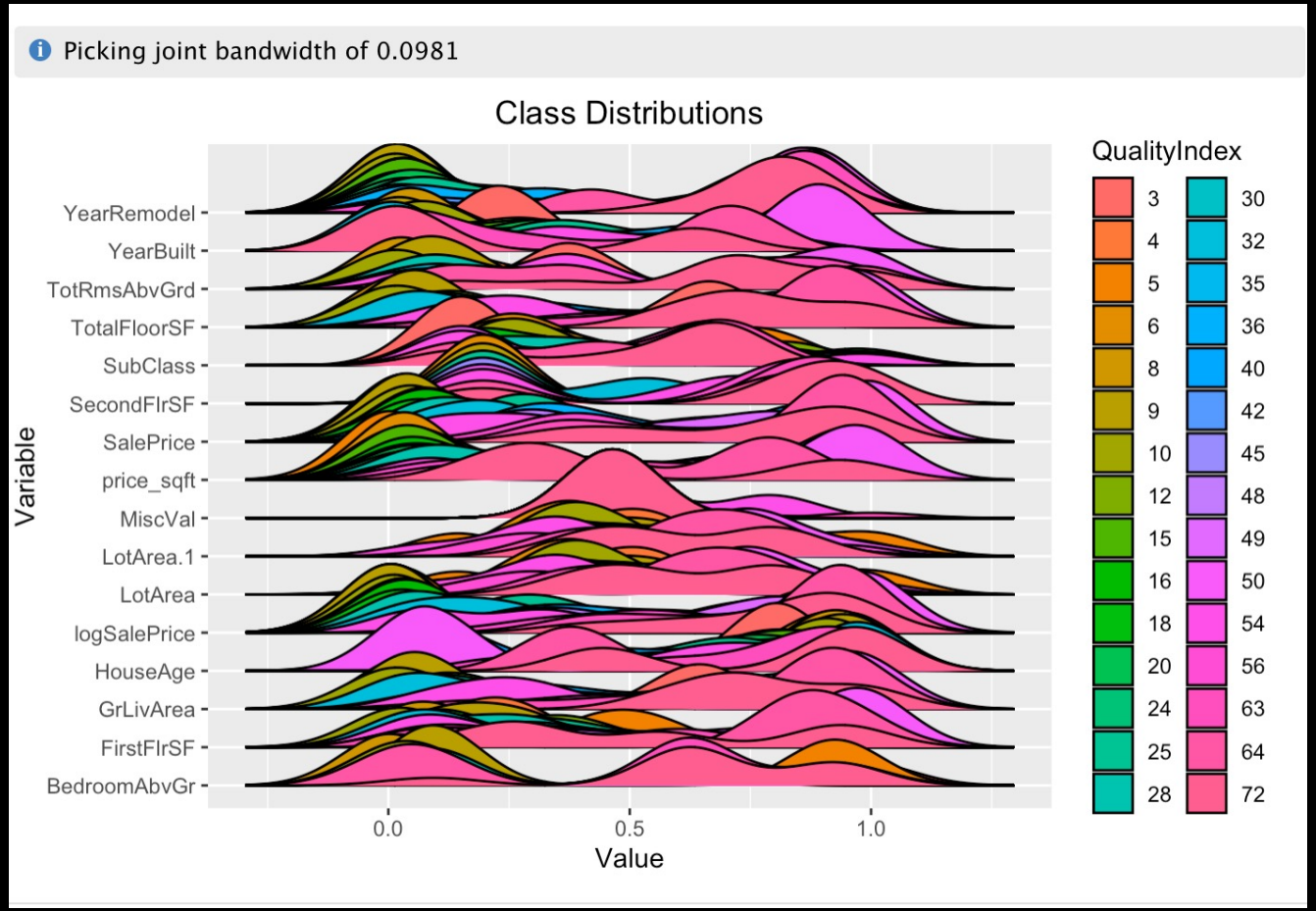


Clusters

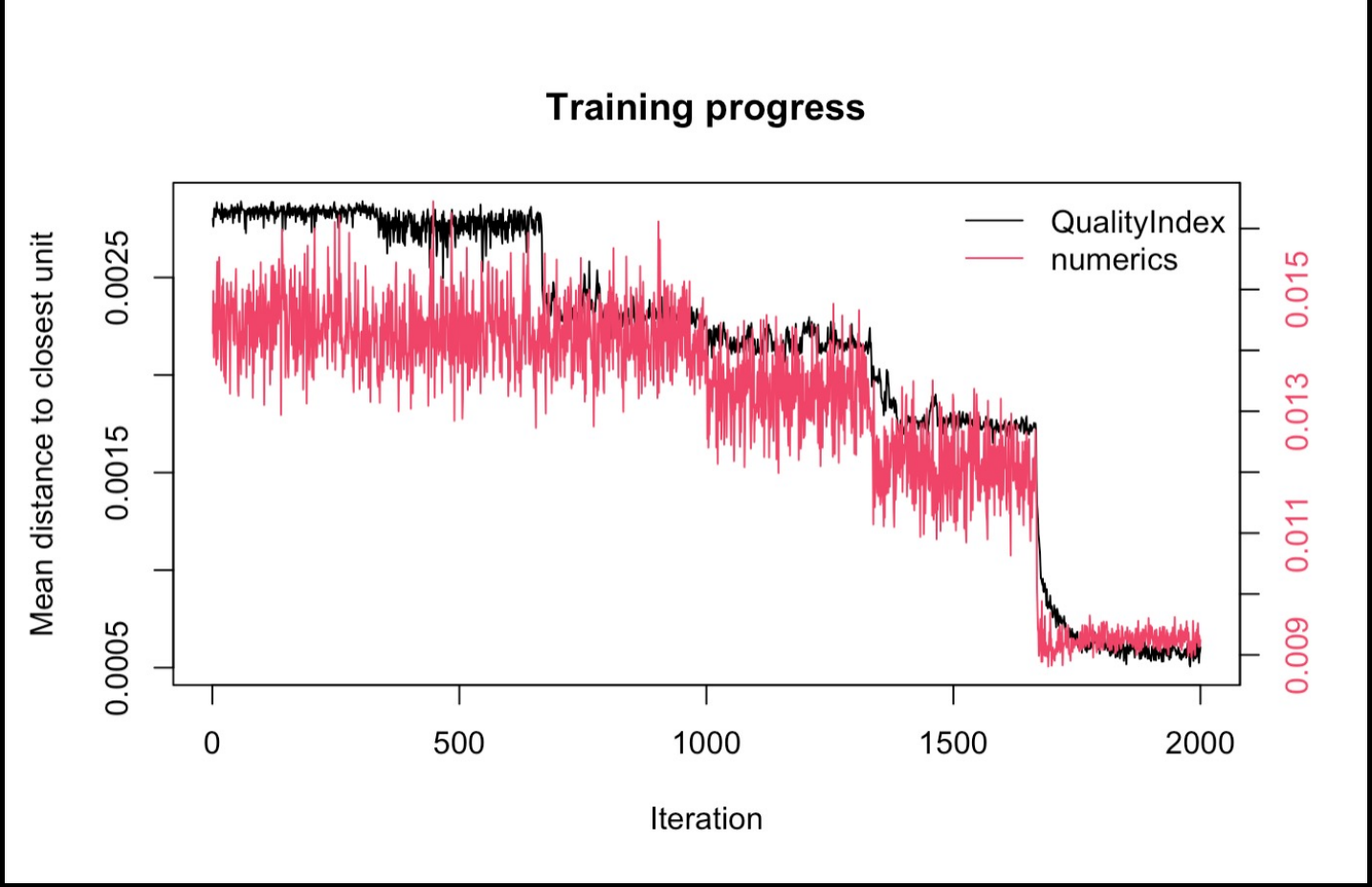


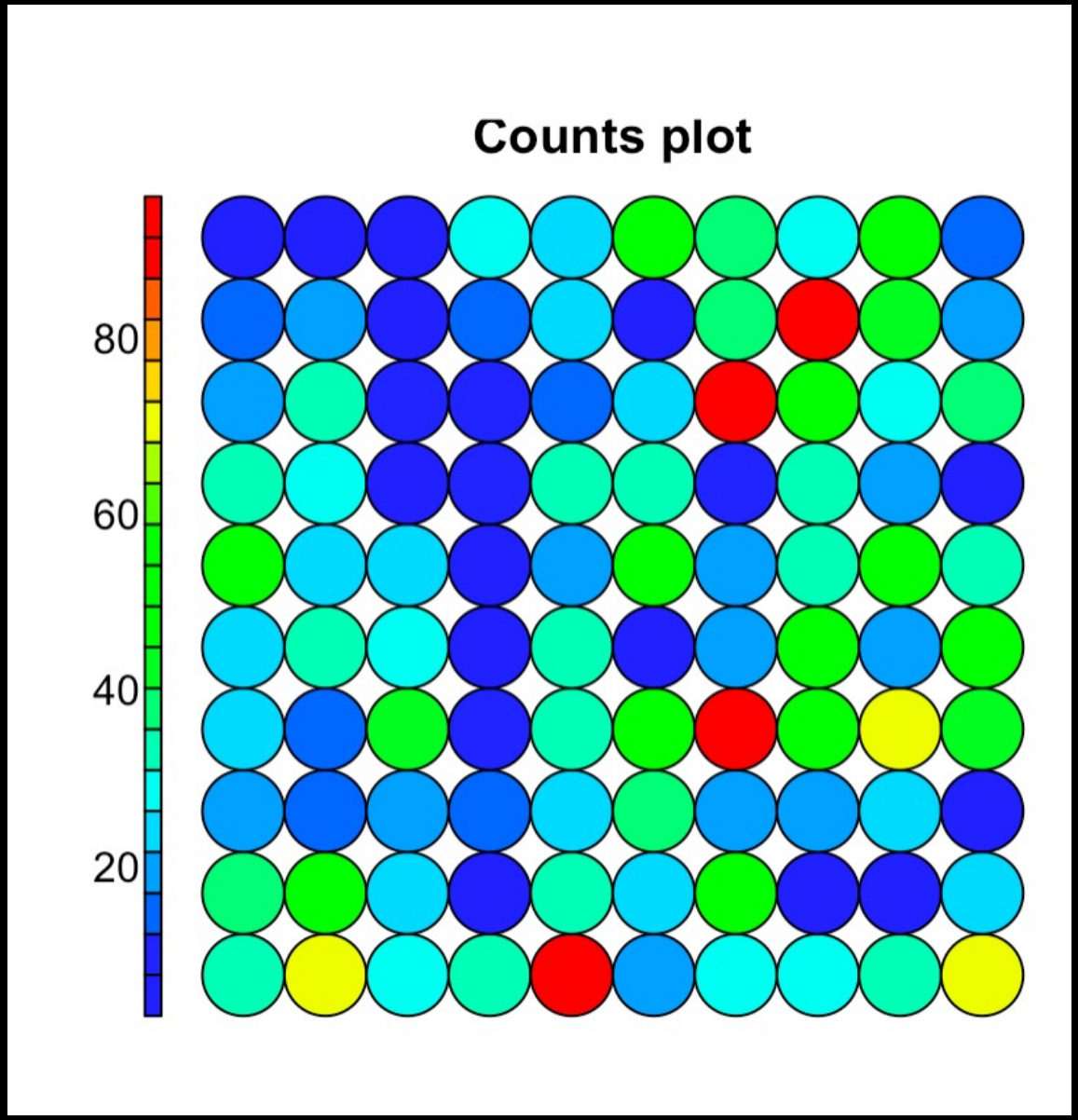
CHECKING FOR ANOMALIES VIA SOM (OVERALL QUALITY)

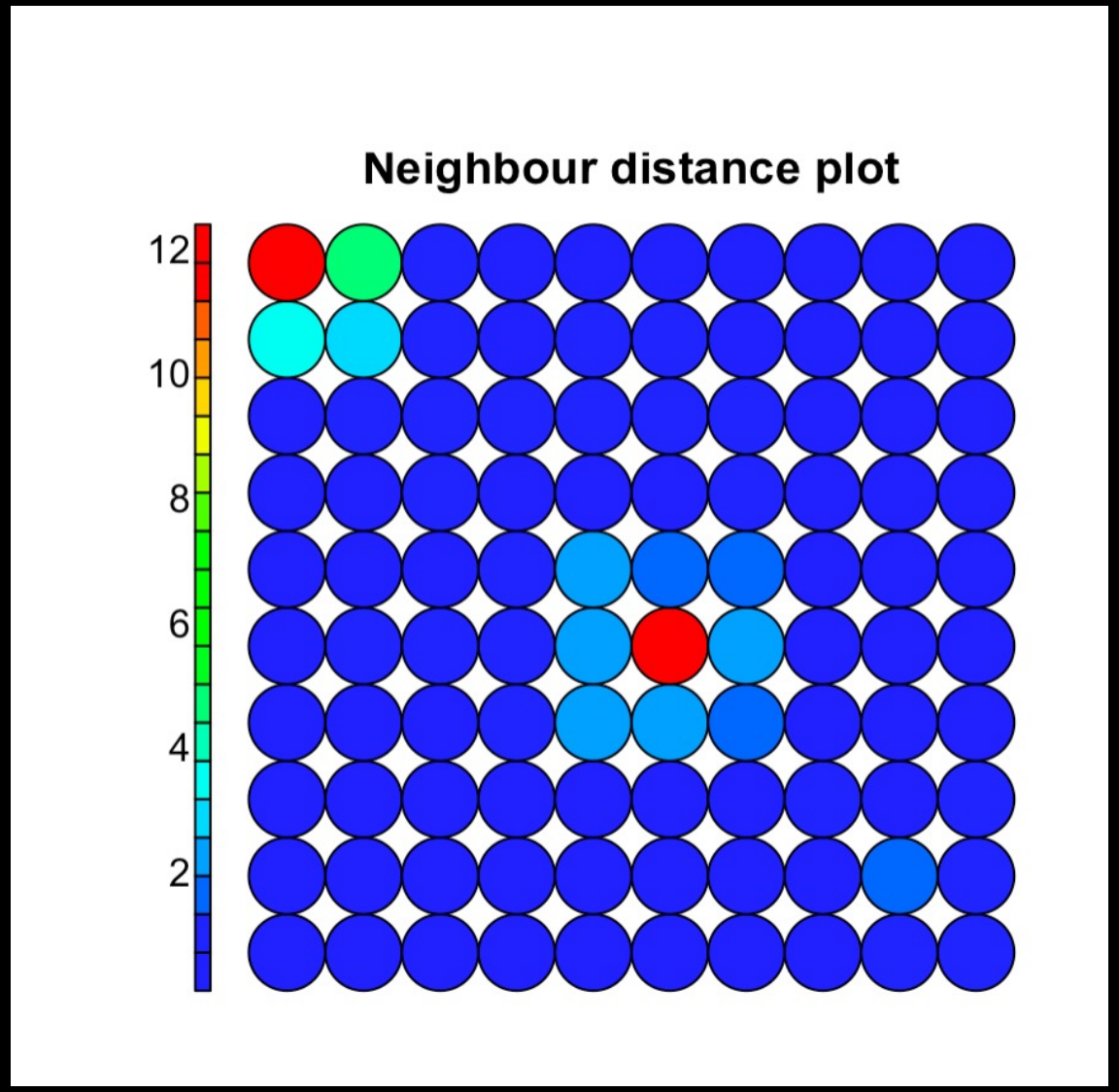
QUALITY INDEX



TRAINING


Training progress
QualityIndex
numerics
Mean distance to closest unit
0.0005
0.0015
0.0025
0.009
0.011
0.013
0.015
0
500
1000
1500
2000
Iteration


Counts plot
80
60
40
20

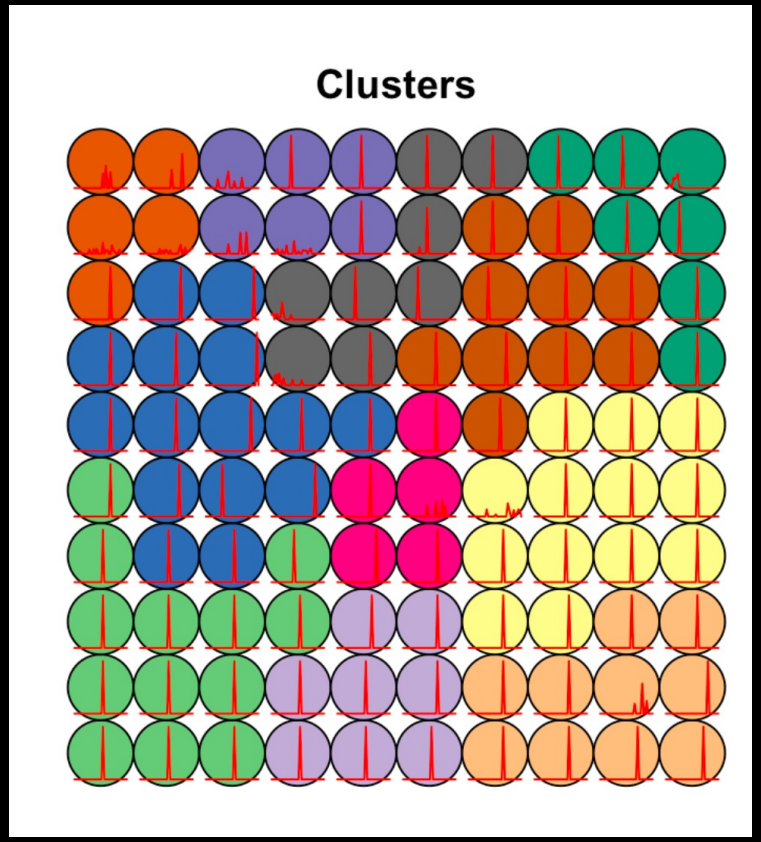

Neighbour distance plot
12
10
8
6
4
2

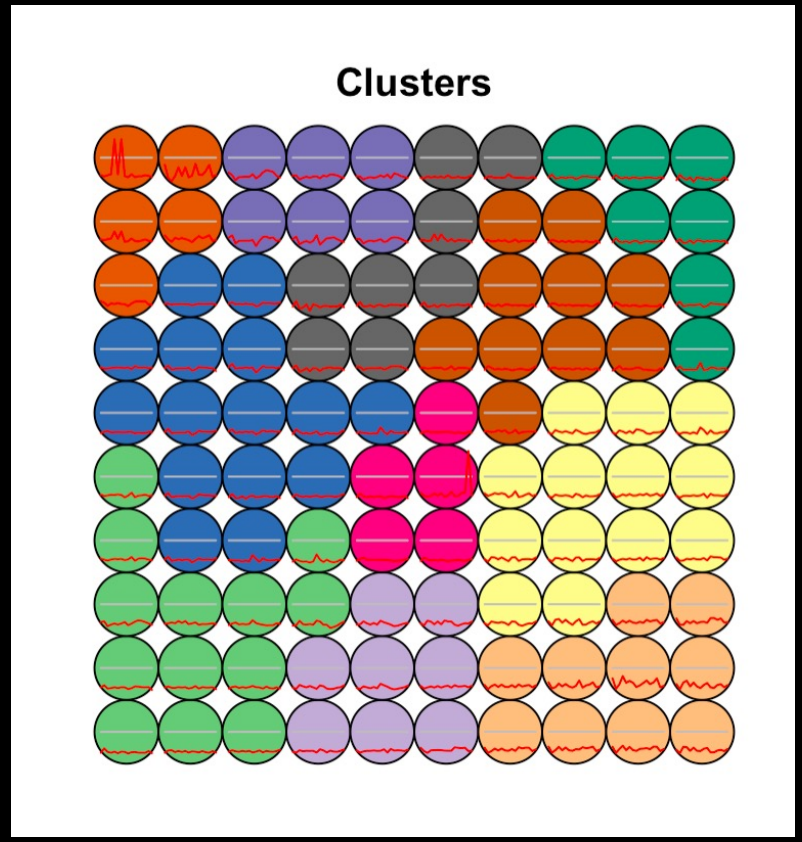
	TotalFloorSF <dbl>	HouseAge <dbl>	QualityIndex <fctr>	price_sqft <dbl>	SalePrice <dbl>	LotArea <dbl>	logSalePrice <dbl>	LotArea.1 <dbl>
39	0.88443584	-1.1697769	45	2.5745833	2.6837504	0.001405854	2.1250886	0.001405854
42	0.40344699	-1.0377260	45	1.2654704	1.1792194	0.174121201	1.2354609	0.174121201
45	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

	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


Clusters


Clusters

LIMITATIONS

- Some limitations for this project :
- 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.
- Lastly, I would have 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 general. I would have loved to go into the specifics.