

Housing Data Analysis

Brooke Fitzgerald, Gloria Giramahoro, Sergi Drago

16/05/2017

Introduction

This data analysis project set out to answer two questions: 1. Which variables are most important to determining the sale price of a house? 2. Are there natural clusters of houses that can be derived and explained from within the data and what do those clusters correspond to?

The dataset we set out to analyze was found on the Kaggle Competition: “House Prices: Advanced Regression Techniques.” The original data comes from the Ames City Assessor’s Office in the form of a data dump from their records system. The initial Excel file contained 113 variables describing 3970 property sales that had occurred in this city between 2006 and 2010. This data was obtained by Prof. De Cock (from Truman State University) who removed variables that required special knowledge or previous calculations for their use, leaving us with the 80 variable dataset we have now.

Loading and cleaning the data

The first step in the data analysis process is data cleaning. This dataset needed quite a bit of cleaning. Many of the variables were only valid in some situations, like PoolQC - a variable that encodes the quality of a pool. In the data descriptions, it stated that if a house didn’t have a pool, the variable was encoded as NA. This also artificially inflated the number of rows that were missing values when in fact the NA was simply coding for the lack of an attribute (pool, garage, deck, etc.).

Therefore, we decided that for all of the variables that used NA to encode a missing attribute, we replaced the NA values for a factor “None” to explicitly encode for that missing attribute. This brought the number of complete observations up from 0 to 1451 out of 1460.

```
## [1] "Initial complete observations:  0"
## [1] "Complete observations coded for missing attributes:  1451"
```

Exploratory Data Analysis

A summary of the data shows that many variables have levels with only a few observations in them. For example, the RoofMatl variable (which describes the material of the roof) has 8 different options, clay tile, standard composite shingle, membrane, metal, roll, gravel and tar, wood shakes, and wood shingles. However, out of all 1460 observations, clay tile, membrane, metal, and roll roofs only have one observation, wood shake has 5 observations, and wood shingle has 6.

A similar pattern is true for many other categorical variables, where certain categories are very sparsely represented. This is potentially a problem for building cross validated models due to the fact that a model that is trained on a subset of the data is unlikely to see many examples of all categories and thus might be unable to accurately model the relationships between the categories and sale price.

##	Id	MSSubClass	MSZoning	LotFrontage
##	Min. : 1.0	Min. : 20.0	C (all): 10	Min. : 0.00
##	1st Qu.: 365.8	1st Qu.: 20.0	FV : 65	1st Qu.: 42.00
##	Median : 730.5	Median : 50.0	RH : 16	Median : 63.00
##	Mean : 730.5	Mean : 56.9	RL : 1151	Mean : 57.62

```

## 3rd Qu.:1095.2 3rd Qu.: 70.0 RM : 218 3rd Qu.: 79.00
## Max. :1460.0 Max. :190.0 Max. :313.00
##
## LotArea Street Alley LotShape LandContour
## Min. : 1300 Grvl: 6 Grvl: 50 IR1:484 Bnk: 63
## 1st Qu.: 7554 Pave:1454 Pave: 41 IR2: 41 HLS: 50
## Median : 9478 None:1369 IR3: 10 Low: 36
## Mean : 10517 Reg:925 Lvl:1311
## 3rd Qu.: 11602
## Max. :215245
##
## Utilities LotConfig LandSlope Neighborhood Condition1
## AllPub:1459 Corner : 263 Gtl:1382 NAmes :225 Norm :1260
## NoSeWa: 1 CulDSac: 94 Mod: 65 CollgCr:150 Feedr : 81
## FR2 : 47 Sev: 13 OldTown:113 Artery : 48
## FR3 : 4 Edwards:100 RRAn : 26
## Inside :1052 Somerst: 86 PosN : 19
## Gilbert: 79 RRAe : 11
## (Other):707 (Other): 15
##
## Condition2 BldgType HouseStyle OverallQual
## Norm :1445 1Fam :1220 1Story :726 Min. : 1.000
## Feedr : 6 2fmCon: 31 2Story :445 1st Qu.: 5.000
## Artery : 2 Duplex: 52 1.5Fin :154 Median : 6.000
## PosN : 2 Twnhs : 43 SLvl : 65 Mean : 6.099
## RRNn : 2 TwnhsE: 114 SFoyer : 37 3rd Qu.: 7.000
## PosA : 1 1.5Unf : 14 Max. :10.000
## (Other): 2 (Other): 19
##
## OverallCond YearBuilt YearRemodAdd RoofStyle
## Min. :1.000 Min. :1872 Min. :1950 Flat : 13
## 1st Qu.:5.000 1st Qu.:1954 1st Qu.:1967 Gable :1141
## Median :5.000 Median :1973 Median :1994 Gambrel: 11
## Mean :5.575 Mean :1971 Mean :1985 Hip : 286
## 3rd Qu.:6.000 3rd Qu.:2000 3rd Qu.:2004 Mansard: 7
## Max. :9.000 Max. :2010 Max. :2010 Shed : 2
##
## RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea
## CompShg:1434 VinylSd:515 VinylSd:504 BrkCmn : 15 Min. : 0.0
## Tar&Grv: 11 HdBoard:222 MetalSd:214 BrkFace:445 1st Qu.: 0.0
## WdShngl: 6 MetalSd:220 HdBoard:207 None :864 Median : 0.0
## WdShake: 5 Wd Sdng:206 Wd Sdng:197 Stone :128 Mean : 103.7
## ClyTile: 1 Plywood:108 Plywood:142 NA's : 8 3rd Qu.: 166.0
## Membran: 1 CemntBd: 61 CmentBd: 60 Max. :1600.0
## (Other): 2 (Other):128 (Other):136 NA's :8
##
## ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
## Ex: 52 Ex: 3 BrkTil:146 Ex :121 Fa : 45 Av :221
## Fa: 14 Fa: 28 CBlock:634 Fa : 35 Gd : 65 Gd :134
## Gd:488 Gd: 146 PConc :647 Gd :618 Po : 2 Mn :114
## TA:906 Po: 1 Slab : 24 TA :649 TA :1311 No :953
## TA:1282 Stone : 6 None: 37 None: 37 None: 38
## Wood : 3
##
## BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2
## ALQ :220 Min. : 0.0 ALQ : 19 Min. : 0.00
## BLQ :148 1st Qu.: 0.0 BLQ : 33 1st Qu.: 0.00

```

```

## GLQ :418      Median : 383.5   GLQ : 14      Median : 0.00
## LwQ : 74      Mean : 443.6     LwQ : 46      Mean : 46.55
## Rec :133      3rd Qu.: 712.2    Rec : 54      3rd Qu.: 0.00
## Unf :430      Max. :5644.0     Unf :1256     Max. :1474.00
## None: 37      None: 38
## BsmUnfSF      TotalBsmSF      Heating      HeatingQC CentralAir
## Min. : 0.0    Min. : 0.0    Floor: 1     Ex:741      N: 95
## 1st Qu.: 223.0 1st Qu.: 795.8 GasA :1428    Fa: 49      Y:1365
## Median : 477.5 Median : 991.5 GasW : 18     Gd:241
## Mean : 567.2   Mean :1057.4 Grav : 7      Po: 1
## 3rd Qu.: 808.0 3rd Qu.:1298.2 OthW : 2     TA:428
## Max. :2336.0   Max. :6110.0 Wall : 4
##
## Electrical      X1stFlrSF      X2ndFlrSF      LowQualFinSF
## FuseA: 94      Min. : 334      Min. : 0        Min. : 0.000
## FuseF: 27      1st Qu.: 882    1st Qu.: 0       1st Qu.: 0.000
## FuseP: 3        Median :1087     Median : 0        Median : 0.000
## Mix : 1         Mean :1163       Mean : 347        Mean : 5.845
## SBrkr:1334      3rd Qu.:1391    3rd Qu.: 728     3rd Qu.: 0.000
## NA's : 1        Max. :4692       Max. :2065        Max. :572.000
##
## GrLivArea      BsmtFullBath      BsmtHalfBath      FullBath
## Min. : 334      Min. :0.0000      Min. :0.00000      Min. :0.000
## 1st Qu.:1130    1st Qu.:0.0000    1st Qu.:0.00000    1st Qu.:1.000
## Median :1464    Median :0.0000    Median :0.00000    Median :2.000
## Mean :1515      Mean :0.4253      Mean :0.05753      Mean :1.565
## 3rd Qu.:1777    3rd Qu.:1.0000    3rd Qu.:0.00000    3rd Qu.:2.000
## Max. :5642      Max. :3.0000      Max. :2.00000      Max. :3.000
##
## HalfBath      BedroomAbvGr      KitchenAbvGr      KitchenQual
## Min. :0.0000    Min. :0.000      Min. :0.000      Ex:100
## 1st Qu.:0.0000  1st Qu.:2.000    1st Qu.:1.000    Fa: 39
## Median :0.0000  Median :3.000    Median :1.000    Gd:586
## Mean :0.3829    Mean :2.866      Mean :1.047      TA:735
## 3rd Qu.:1.0000  3rd Qu.:3.000    3rd Qu.:1.000
## Max. :2.0000    Max. :8.000      Max. :3.000
##
## TotRmsAbvGrd      Functional      Fireplaces      FireplaceQu      GarageType
## Min. : 2.000      Maj1: 14      Min. :0.000      Ex : 24      2Types : 6
## 1st Qu.: 5.000      Maj2: 5       1st Qu.:0.000      Fa : 33      Attchd :870
## Median : 6.000      Min1: 31      Median :1.000      Gd :380      Basement: 19
## Mean : 6.518      Min2: 34      Mean :0.613      Po : 20      BuiltIn: 88
## 3rd Qu.: 7.000      Mod : 15      3rd Qu.:1.000      TA :313      CarPort: 9
## Max. :14.000      Sev : 1       Max. :3.000      None:690     Detchd :387
## Typ :1360      None : 81
## GarageFinish      GarageCars      GarageArea      GarageQual      GarageCond
## Fin :352      Min. :0.000      Min. : 0.0      Ex : 3      Ex : 2
## RFn :422      1st Qu.:1.000    1st Qu.: 334.5    Fa : 48      Fa : 35
## Unf :605      Median :2.000    Median : 480.0    Gd : 14      Gd : 9
## None: 81      Mean :1.767      Mean : 473.0     Po : 3      Po : 7
## 3rd Qu.:2.000    3rd Qu.: 576.0    TA :1311      TA :1326
## Max. :4.000      Max. :1418.0     None: 81      None: 81
##
## PavedDrive      WoodDeckSF      OpenPorchSF      EnclosedPorch

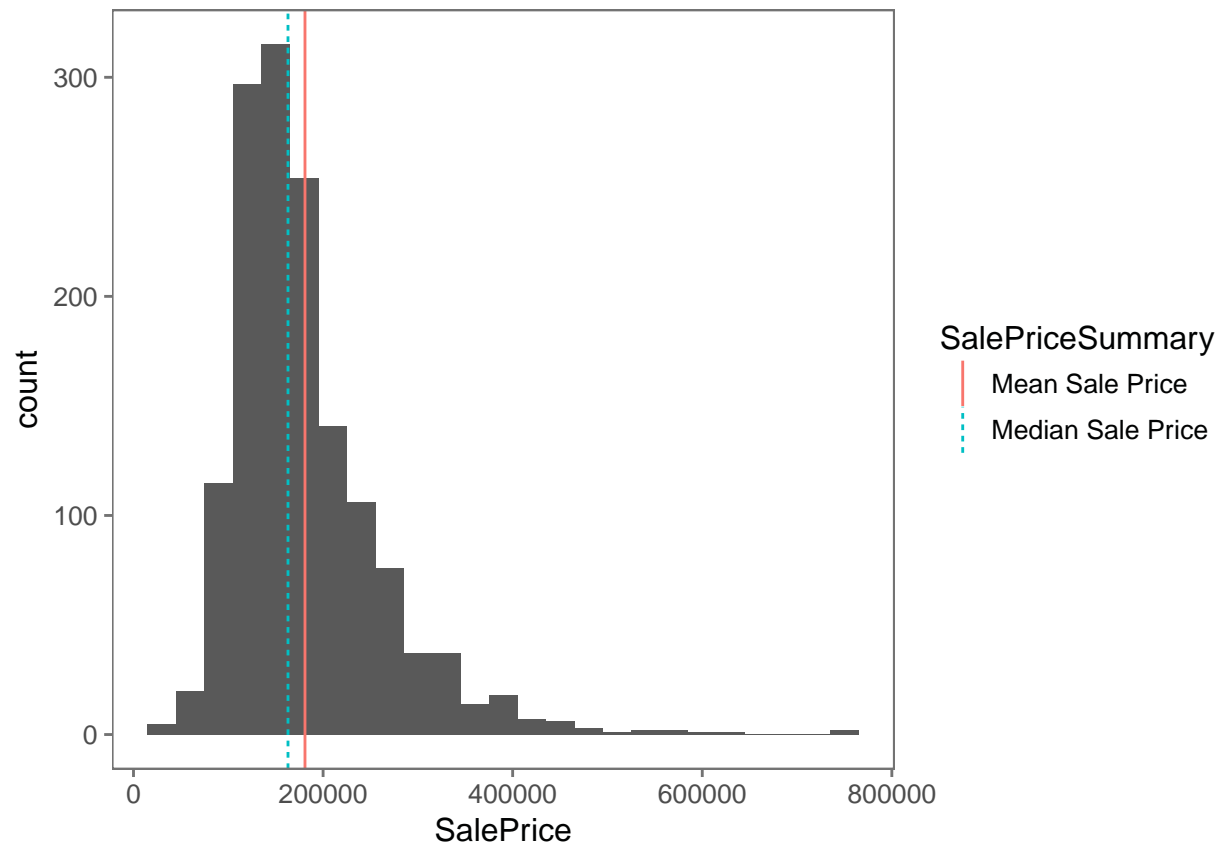
```

```

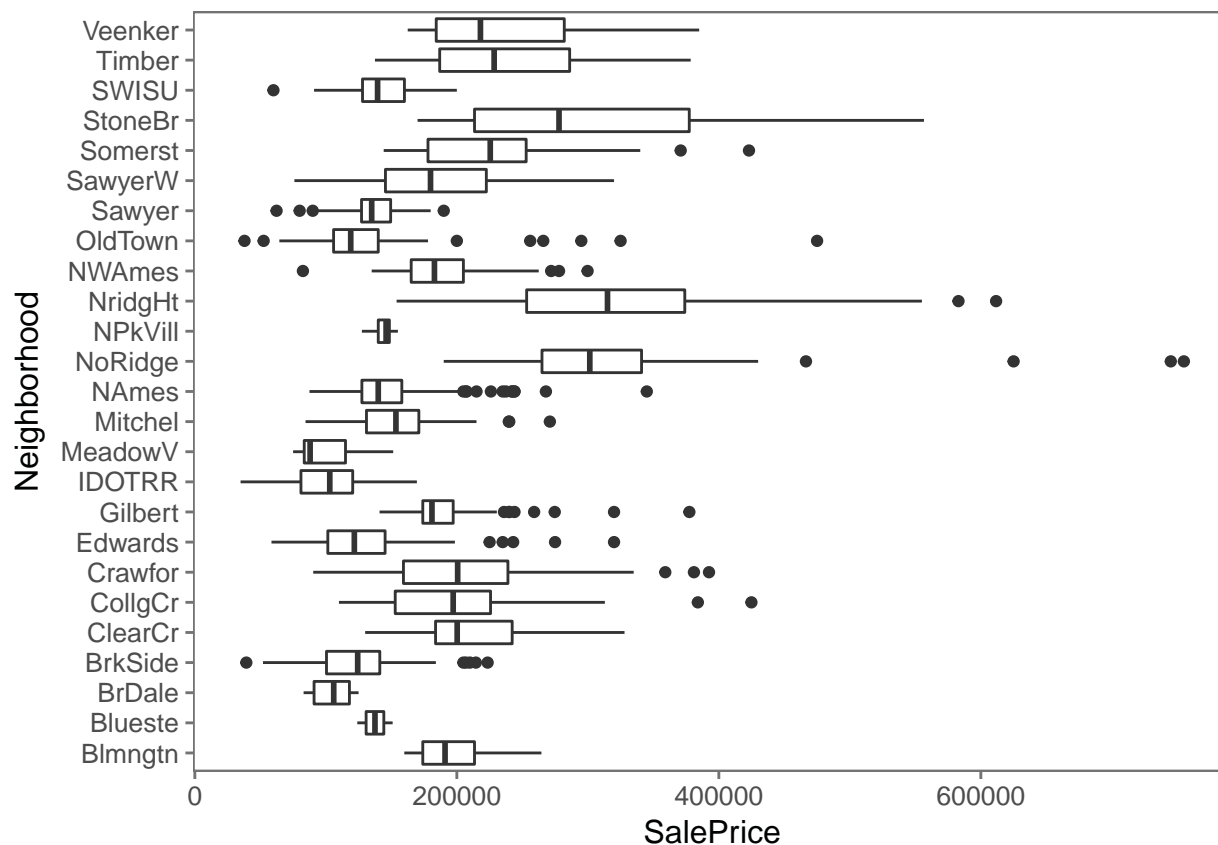
## N: 90      Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
## P: 30      1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00
## Y:1340     Median : 0.00   Median : 25.00   Median : 0.00
##           Mean    : 94.24   Mean    : 46.66   Mean    : 21.95
##           3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.: 0.00
##           Max.    :857.00   Max.    :547.00   Max.    :552.00
##
##      X3SsnPorch      ScreenPorch      PoolArea      PoolQC
## Min.   : 0.00   Min.   : 0.00   Min.   : 0.000   Ex   : 2
## 1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.000   Fa   : 2
## Median : 0.00   Median : 0.00   Median : 0.000   Gd   : 3
## Mean    : 3.41   Mean    :15.06   Mean    : 2.759   None:1453
## 3rd Qu.: 0.00   3rd Qu.: 0.00   3rd Qu.: 0.000
## Max.    :508.00   Max.    :480.00   Max.    :738.000
##
##      Fence      MiscFeature      MiscVal      MoSold
## GdPrv: 59   Gar2: 2   Min.   : 0.00   Min.   : 1.000
## GdWo : 54   Othr: 2   1st Qu.: 0.00   1st Qu.: 5.000
## MnPrv:157   Shed: 49   Median : 0.00   Median : 6.000
## MnWw : 11   TenC: 1   Mean    : 43.49   Mean    : 6.322
## None :1179   None:1406   3rd Qu.: 0.00   3rd Qu.: 8.000
##                                     Max.    :15500.00   Max.    :12.000
##
##      YrSold      SaleType      SaleCondition      SalePrice
## Min.   :2006   WD      :1267   Abnorml: 101   Min.   : 34900
## 1st Qu.:2007   New      : 122   AdjLand: 4     1st Qu.:129975
## Median :2008   COD      : 43   Alloca : 12    Median :163000
## Mean    :2008   ConLD    : 9    Family : 20    Mean    :180921
## 3rd Qu.:2009   ConLI    : 5    Normal :1198   3rd Qu.:214000
## Max.    :2010   ConLw    : 5    Partial: 125   Max.    :755000
##                                     (Other): 9

```

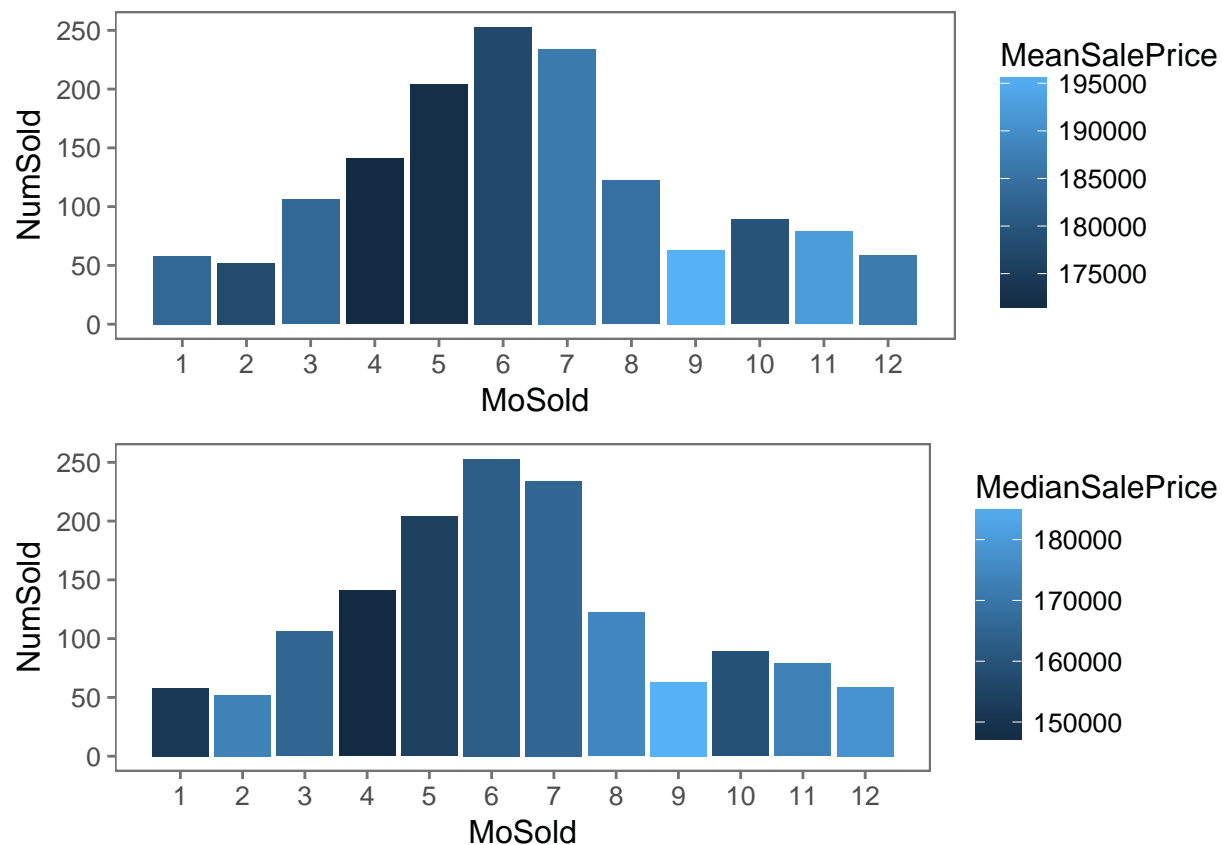
Then with some data visualization we find that the sale price of homes has a positively skewed distribution, with the median sale price of a house (\$163,000.00) being significantly less than the average sale price of a house (\$180,921.20).



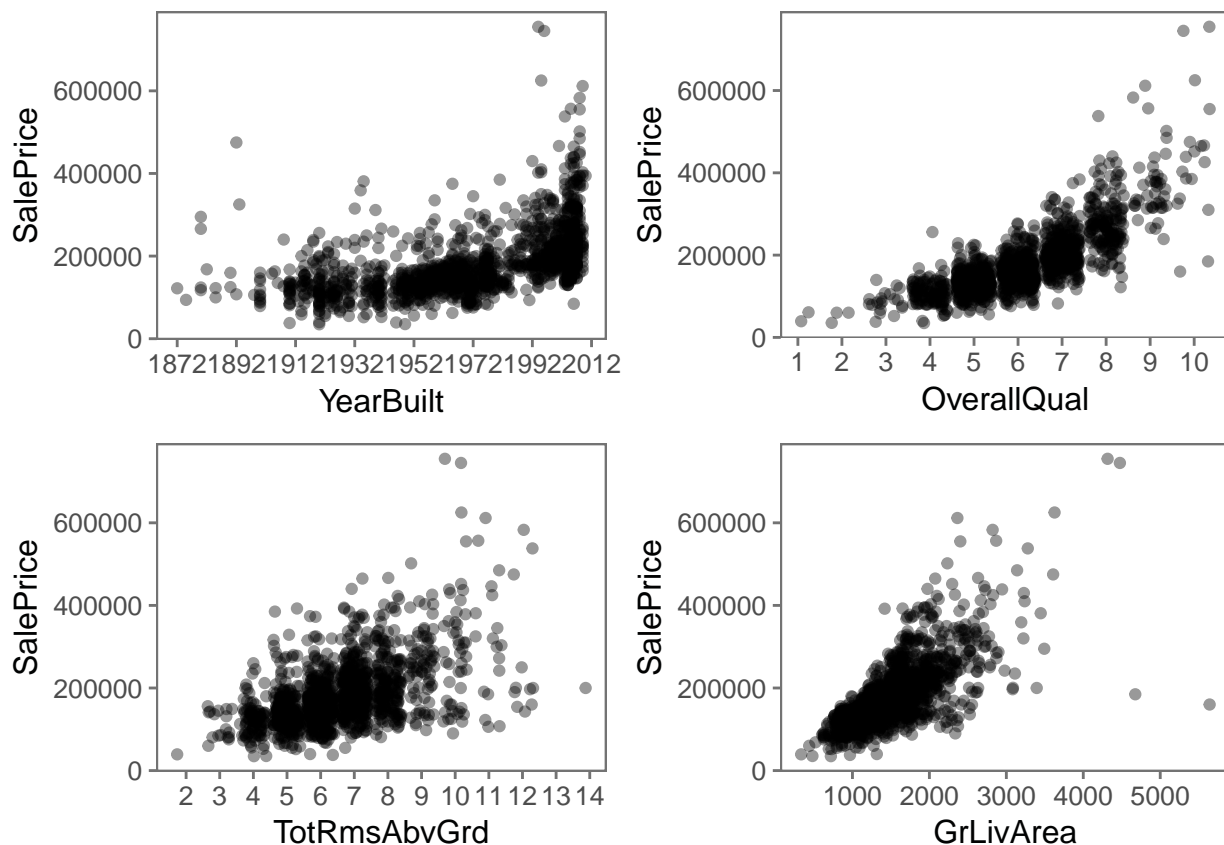
A boxplot comparing the sale prices for all of the different neighborhoods shows that there are some neighborhoods that clearly have much higher average and median sale prices than others. This indicates that Neighborhood is likely an important indicator for sale price.



We also wanted to see if which month a house was purchased in has any effect on the sale price of the house. Based on the plots of the number of sales per month, colored by the mean and median sale price for each month, it seems like house sale prices are lower and more frequent as the summer starts (April, May), but that buying a house in September is going to be much more expensive then during other months.



In the plots of sale price and the year built (YearBuilt), overall house quality (OverallQual), total number of rooms above ground (TotRmsAbvGrd), and above grade living area square feet (GrLivArea), it becomes clear that these three variables are positively correlated with sale price.

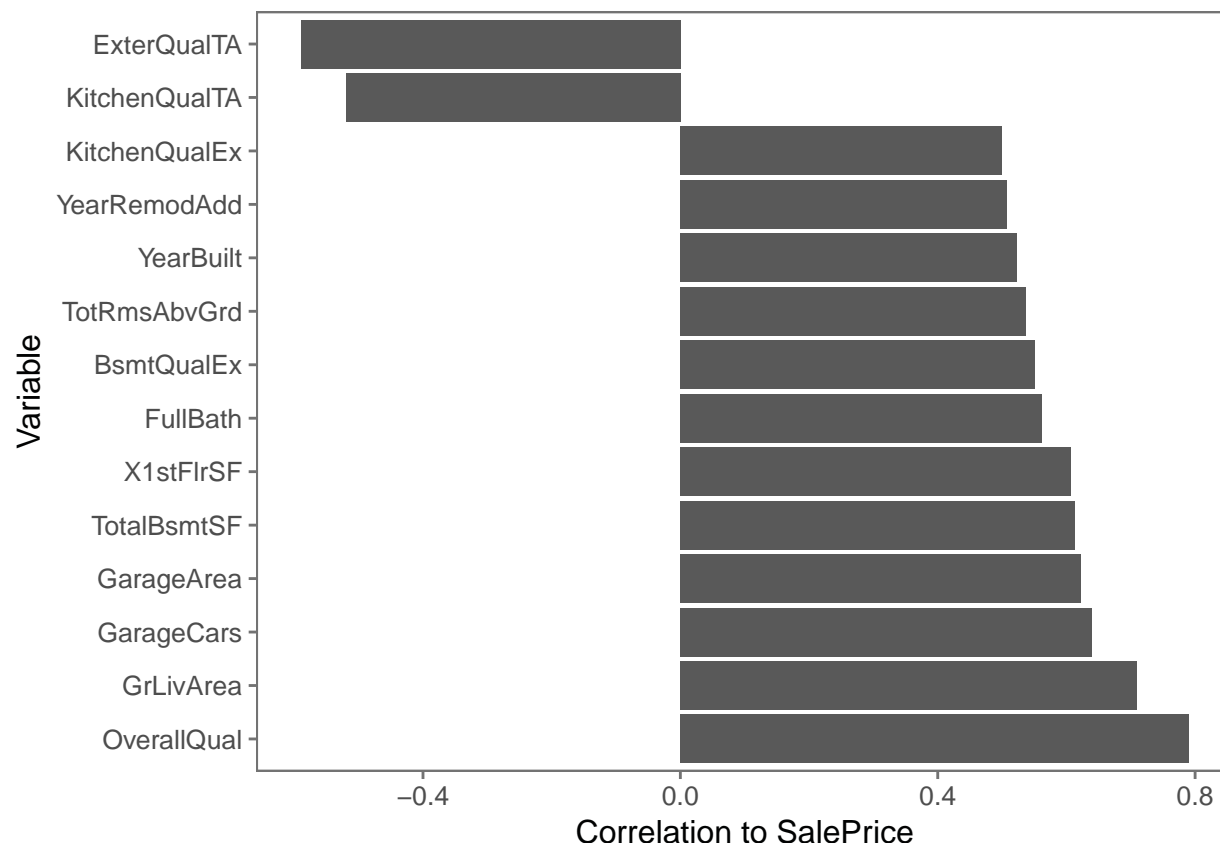


Because year built, overall house quality, total number of rooms above ground, and above grade living area square feet seemed correlated with SalePrice, we wanted to definitively test said correlation and see if there are other variables that are highly correlated with SalePrice.

However, since correlation requires numeric variables, and since many of our variables are factors, we decided to use a technique known as one hot variable encoding. With this technique, each variable with n factors is expanded into n variables that take the value 1 if the data has that factor and 0 otherwise.

For example, the variable Street with its two possible factors 'Gravl' and 'Pave' becomes two variables: StreetGravl and StreetPave that are either 0 or 1.

With this new encoding we calculated the correlation and then plotted the variables with absolute correlation values larger than 0.5.



Removing NA values

Because some predictive models don't allow for NA values, we want to do some data imputation. The data was imputed using random forests. This imputed data can be used in conjunction with the one-hot encoded data to compare and contrast the various predictions.

Starting Predictive Modeling

The first step in predictive modeling is splitting the data into training and test sets.

We did this by using the `createDataPartition` function from the `caret` package in order to get training and testing sets that have values of `SalePrice` that are evenly split across the dataset.

We created training and testing sets for the full data, the imputed data, the one hot encoded data, and the one hot encoded data with NA values removed.

First Research Question

The first model we tried out was a simple decision tree. We created decision trees on all of our training sets and calculated the cross-validated RMSE for each. We also averaged the predictions from each decision tree and calculated the cross-validated RMSE for that. Overall, the predictions from the decision trees weren't too inaccurate, though they tended to underpredict sale price for more expensive homes. Furthermore, the

average of the predictions from all three trees has the lowest RMSE, suggesting that a random forest might increase the accuracy of the model.

We then decided to answer our second research question: Which variables are most important to determining the sale price of a house?

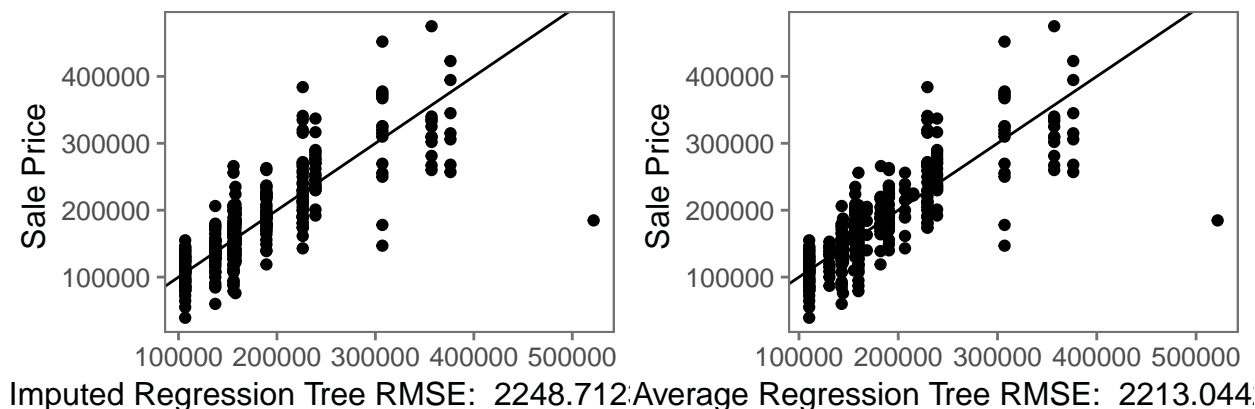
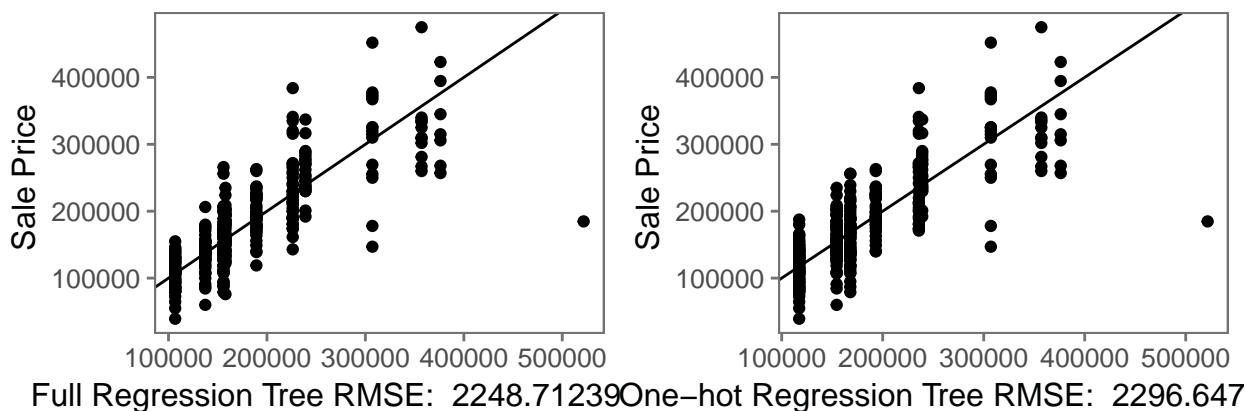
To do this, we inspected the individual decision trees the algorithm constructed. Each algorithm constructed a decision tree that has 11 leaf nodes of sale price bins that are reached with either 3 or 4 splits.

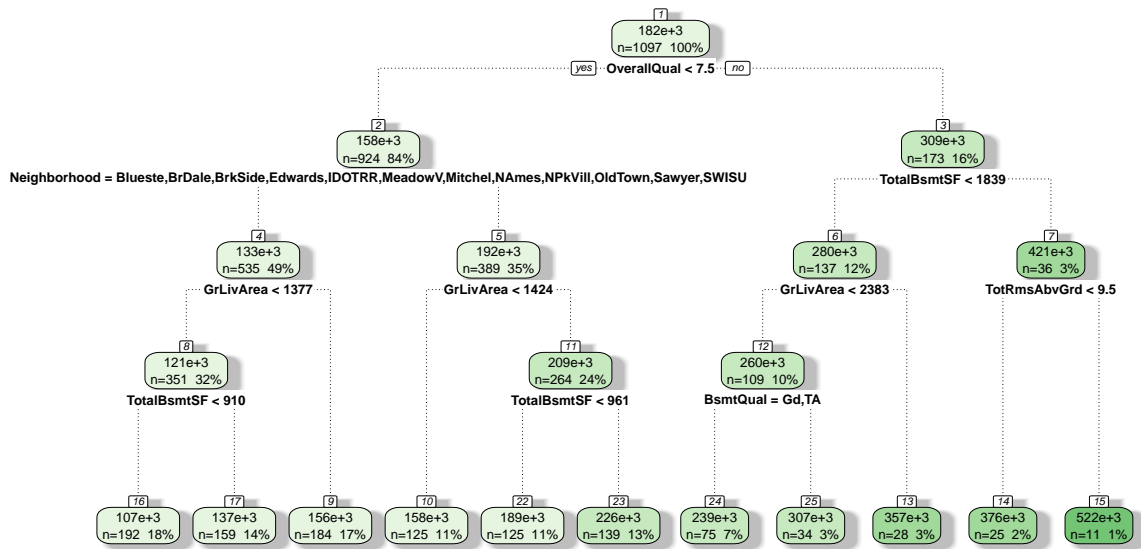
From these, it becomes immediately obvious that the overall quality of a house is the most important factor to sale price, as it was used for all three trees for the first split. After that, the trees are not as synonymous, though they do contain similar information about the important variables in predicting sale price.

All three models use the above grade living area square feet, total basement square feet, total rooms above ground, and basement quality to make further splits. The only variable not used by the one-hot encoded model is the neighborhood, although it is used as a splitting variable in level two of the full and imputed data.

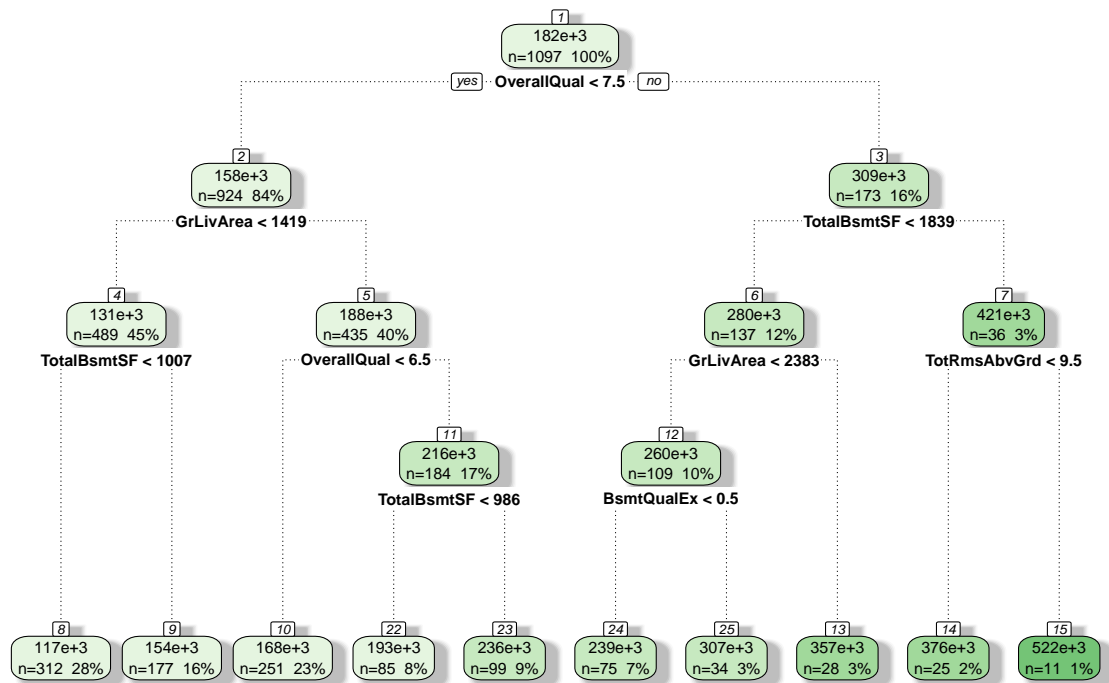
For the full data and the imputed data the input data point is assigned to the lowest three nodes if it is in any of the neighborhoods Briardale, Bluestem, Brookside, Edwards, Iowa DOT and Rail Road, Meadow Village, Mitchell, Old Town, Sawyer, or South & West of Iowa State University. However, for the one-hot encoded data the input data point is assigned to the lowest 3 sale price nodes if the above grade living area square feet is less than 1419 ft^2 .

When you look at the box plots comparing the two splits, the neighborhood split has a smaller interquartile range within sale price than the above grade living area square feet split. This may explain why the one-hot encoded decision tree has a slightly worse RMSE than the other two models.

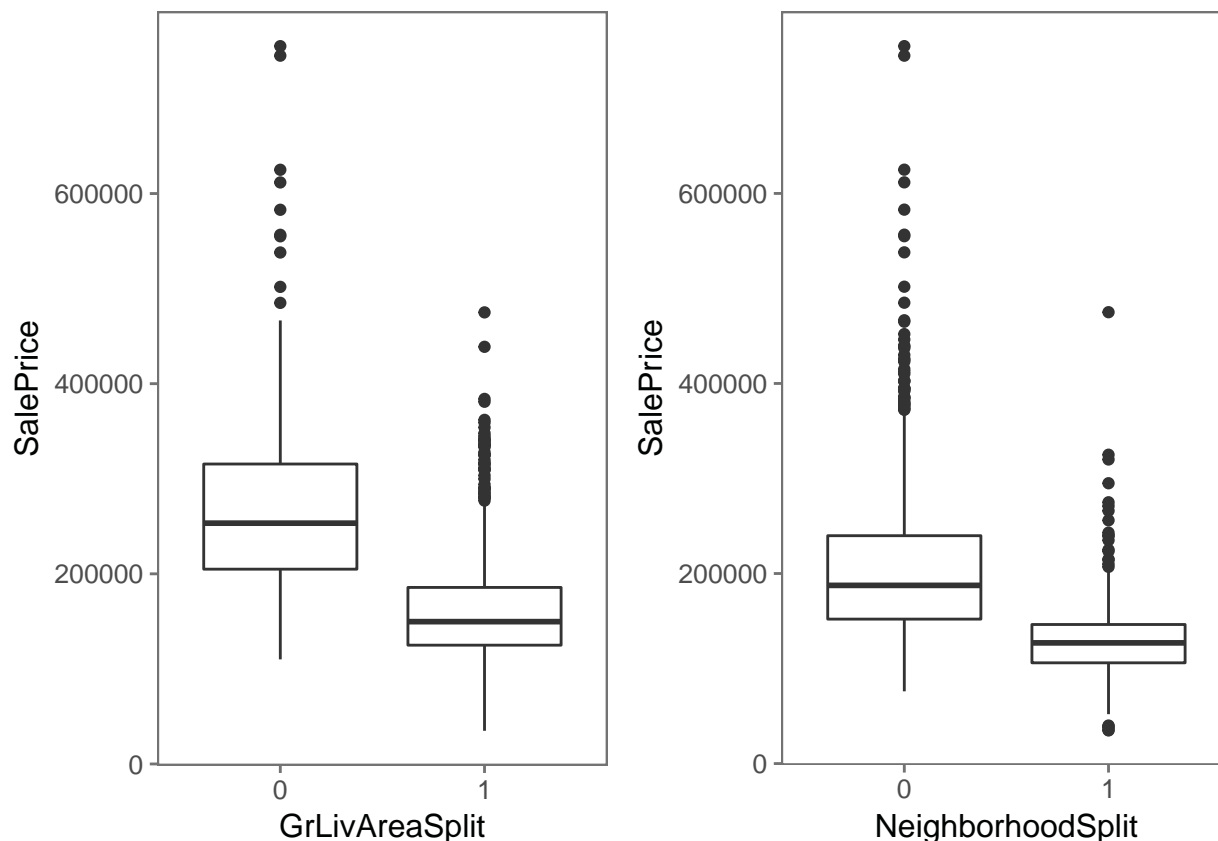




Rattle 2017-May-16 16:14:17 baf14



Rattle 2017-May-16 16:14:18 baf14



The next method we used to attempt to answer our research question is linear regression. Due to some of the factor variables having extremely sparse categories (as discussed above) when the data was split into training and testing sets the testing set occasionally had factors that were missing from the training set. To overcome this, we manually added the factors that were missing, but this understandably caused some deficiencies in the model and its predictions.

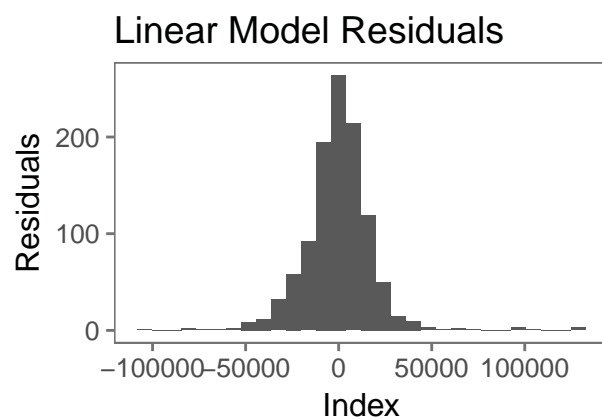
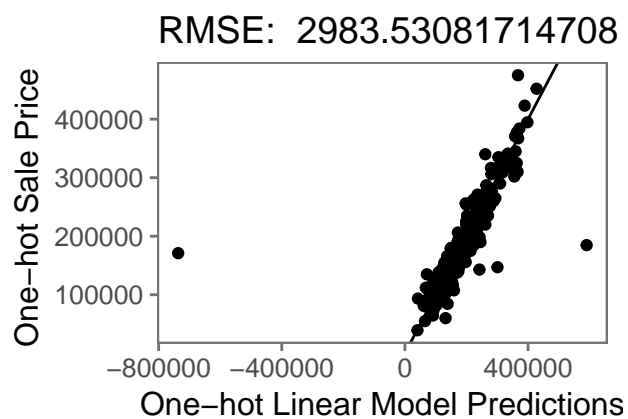
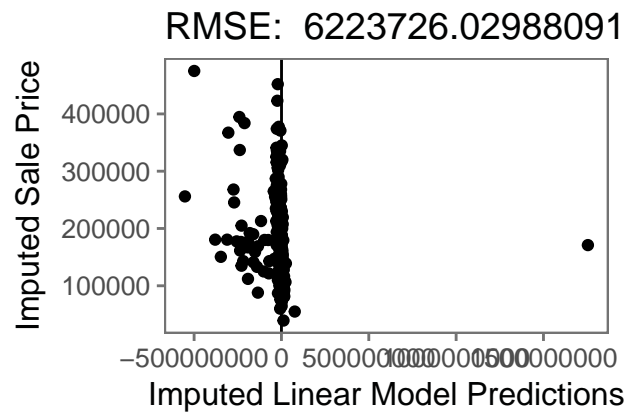
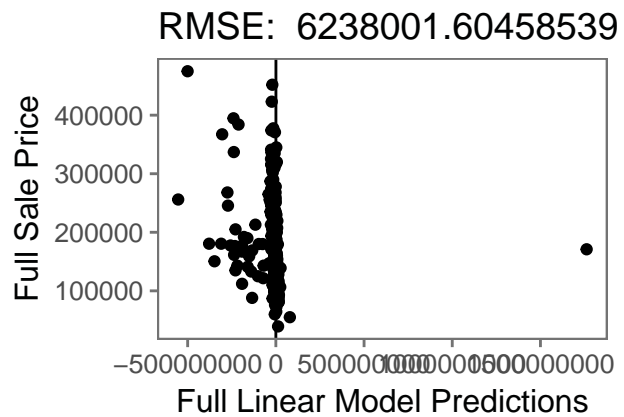
The linear models fit on the full dataset and the imputed dataset entirely failed to predict sale price, likely due to the lack of factor representation. However, the one-hot encoded data will always represent every factor category as a variable, even if that variable contains only zeros in the training split. The predictions from the one hot encoded data fit the data rather well aside from one outlier with an adjusted R^2 of 0.8874 and RMSE of 2983.531.

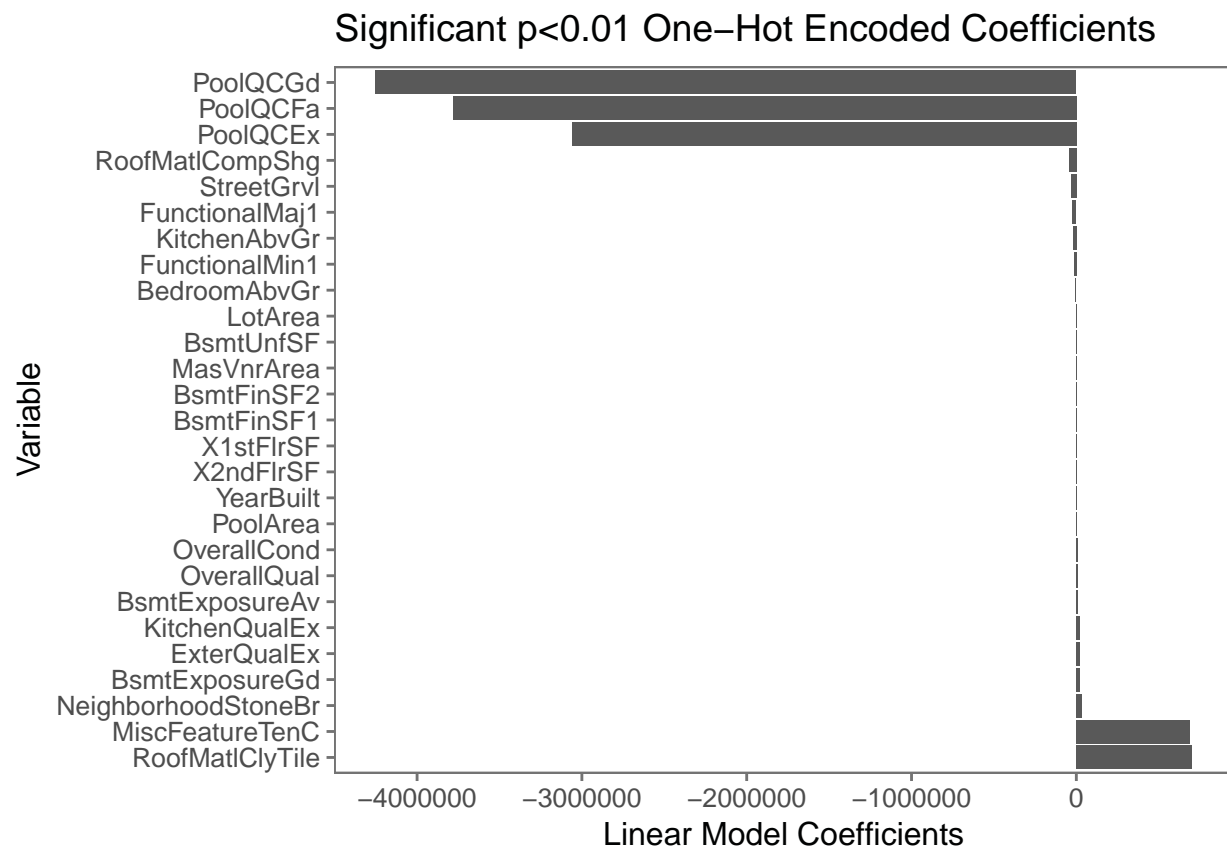
However, when examining the statistically significant coefficients of the model in order to answer our research question, it becomes clear that interpreting the outputs of a model with such sparsely represented factors gives nonsensical results - e.g. that houses with a good quality pool are worth \$4,000,000 less than those without.

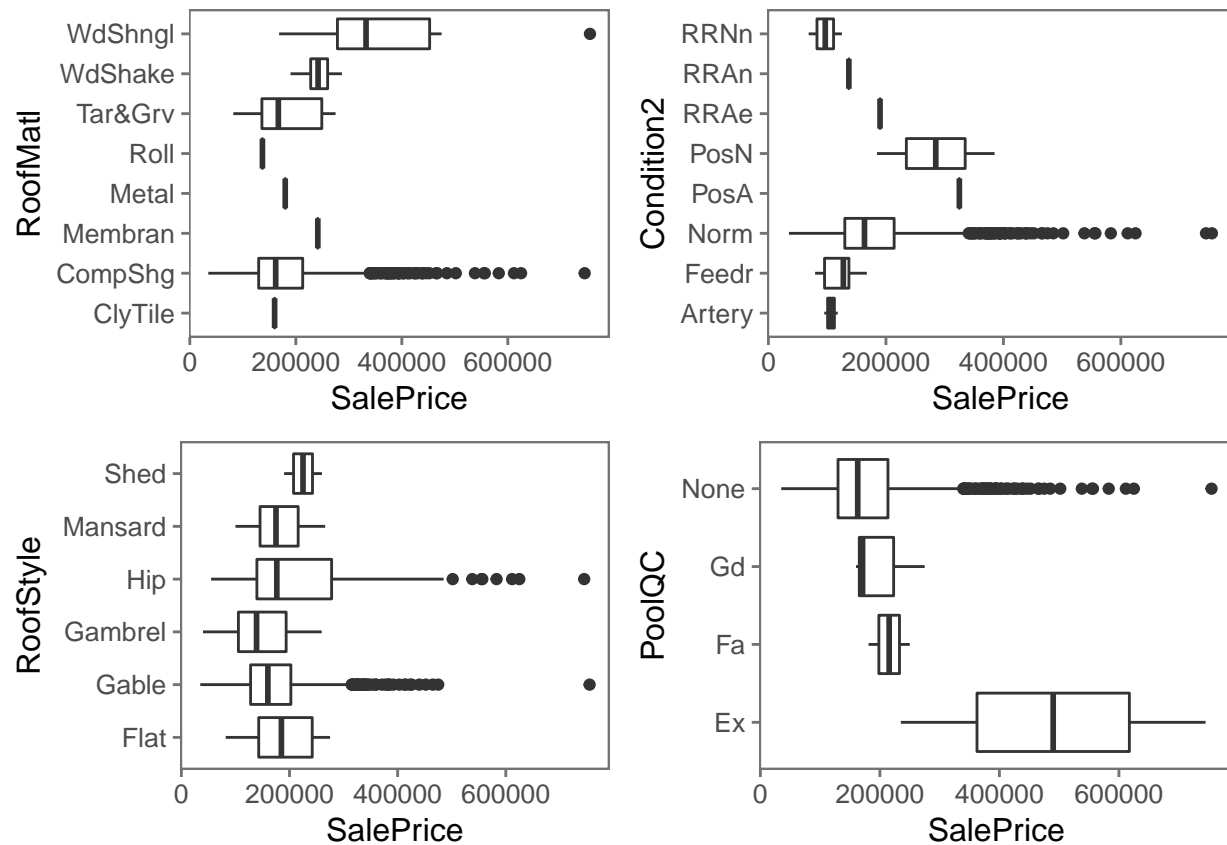
Therefore, we filtered the one-hot encoded data to only get the variables with more than 15 observations in the training set and trained a linear model on that. The RMSE is much lower than the previous model at 1700.314, and the adjusted R^2 is slightly lower at 0.8874, and the coefficients are much more interpretable.

From this linear model, we can say that living in a less nice neighborhood (Mitchell, Northwest Ames, Edwards) having a quick and significant rise from street grade to building has the most significant negative impact on the sale price of a house, while being from a nice neighborhood (Northridge, Northridge Heights, Stone Brook), having excellent kitchen quality, non-irregular lot shape, good quality walkout or garden level walls (BsmtExposureGd), and garage capacity in number in cars all have a large impact on sale price.

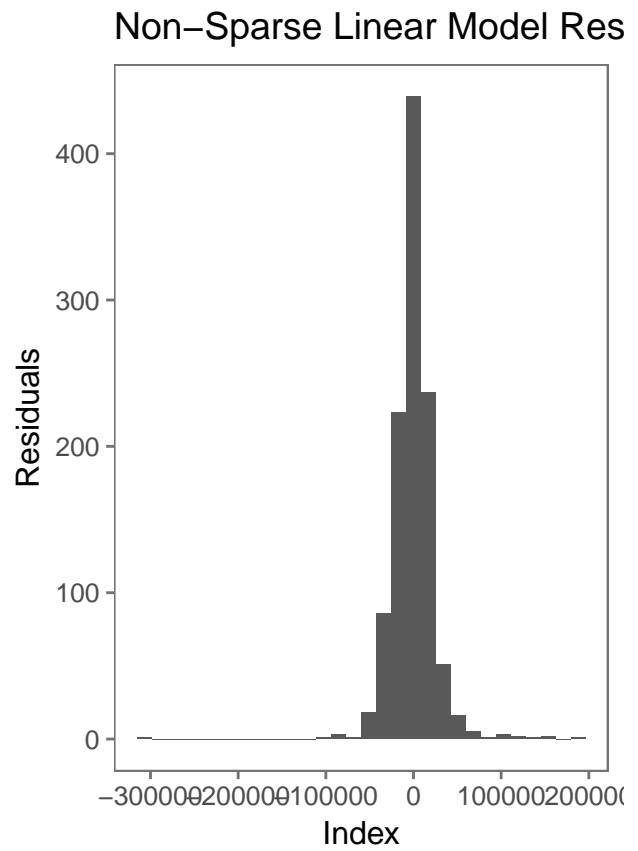
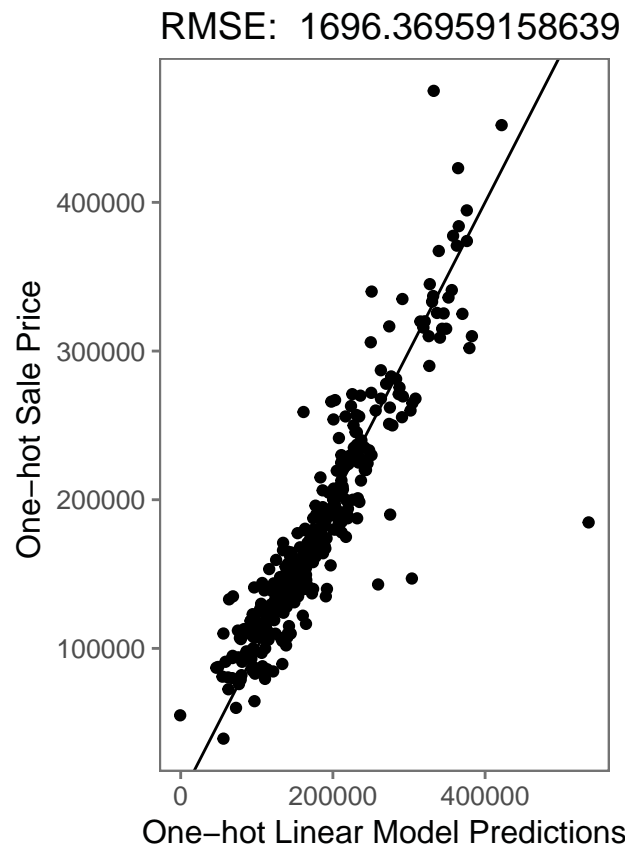
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

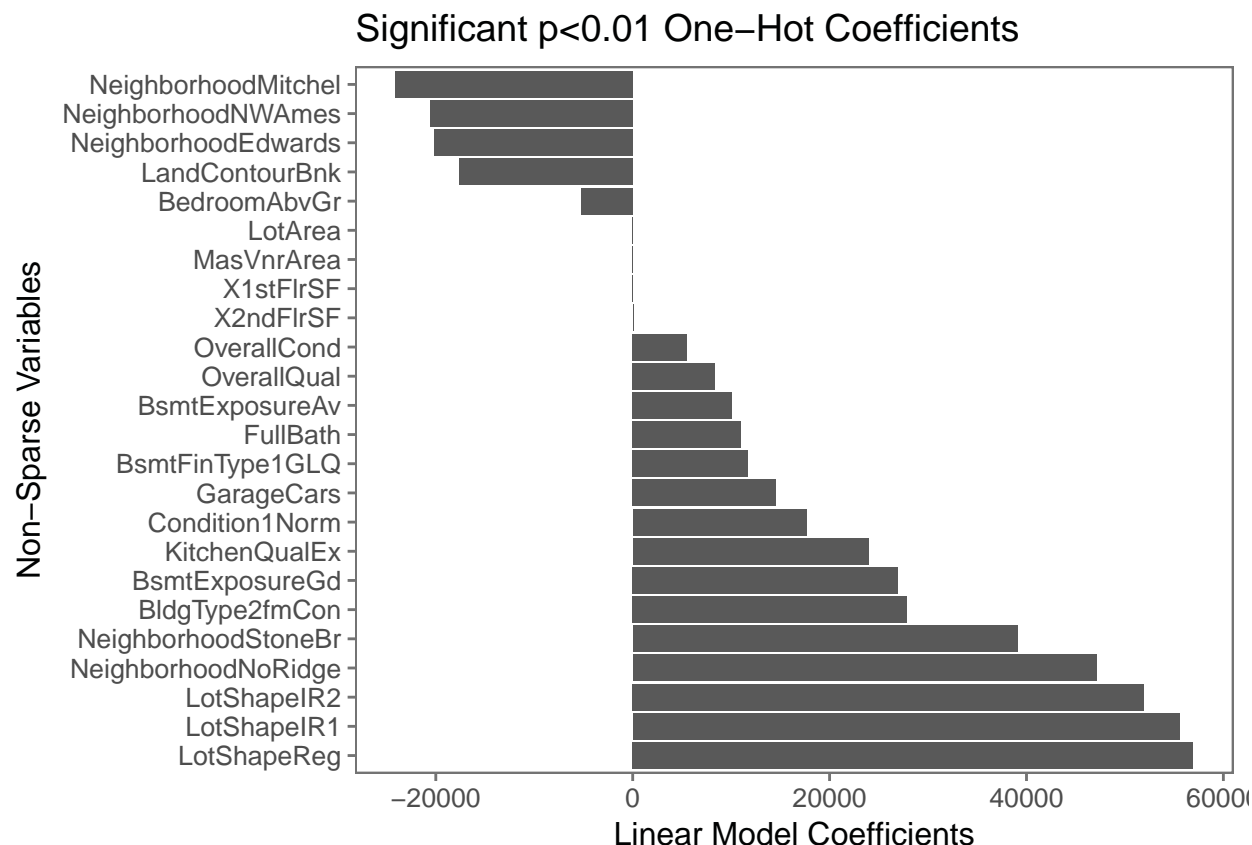






`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.





Forward and backward subset selection was also used to shed some insight on our first research question. Both forward and backward selection selected the house id, overall condition, year remodel added, and fair basement quality and garage area as top five predictive variables. Where they differ is that the forward subset selection seems to select variables concerning the zoning and basement full baths of the house, while the backward subset selection seems to select variables relating to the square feet of the house.

These selected variables should be taken with a grain of salt because the same non-sparse variable selection process was used as with the linear models, but forward and backward subset selection identify variables that weren't identified as important by any other method. Because there are so many variables, it is too computationally intensive to calculate the best subset selection, but it is likely that forward and backward subset selection with this many variables is identifying a non-optimal solution.

```
## Reordering variables and trying again:
```

```
## Reordering variables and trying again:
```

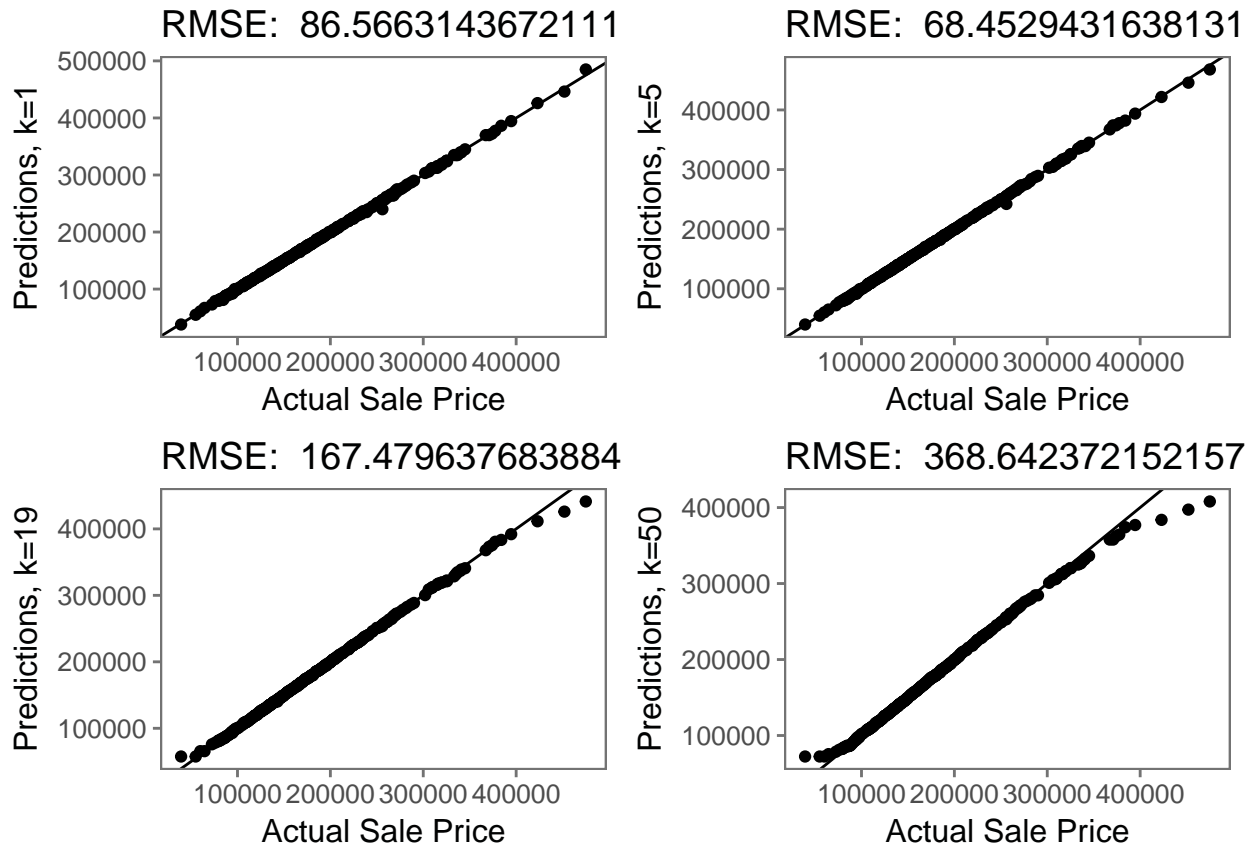
```
## Forward.Selected.Variables Backward.Selected.Variables
## 1 Id Id
## 2 MSZoningFV OverallCond
## 3 OverallCond YearRemodAdd
## 4 YearRemodAdd BsmtQualFa
## 5 BsmtQualFa X2ndFlrSF
## 6 BsmtFullBath LowQualFinSF
```

Below we applied the KNN regression on the SalePrice. The regression is basically fitting the best line to predict the SalePrice between neighbouring houses. We initially assumed 4 values of k: 1,5,19 and 50. Among the 4 values, 5 had the least root mean square error(RMSE).

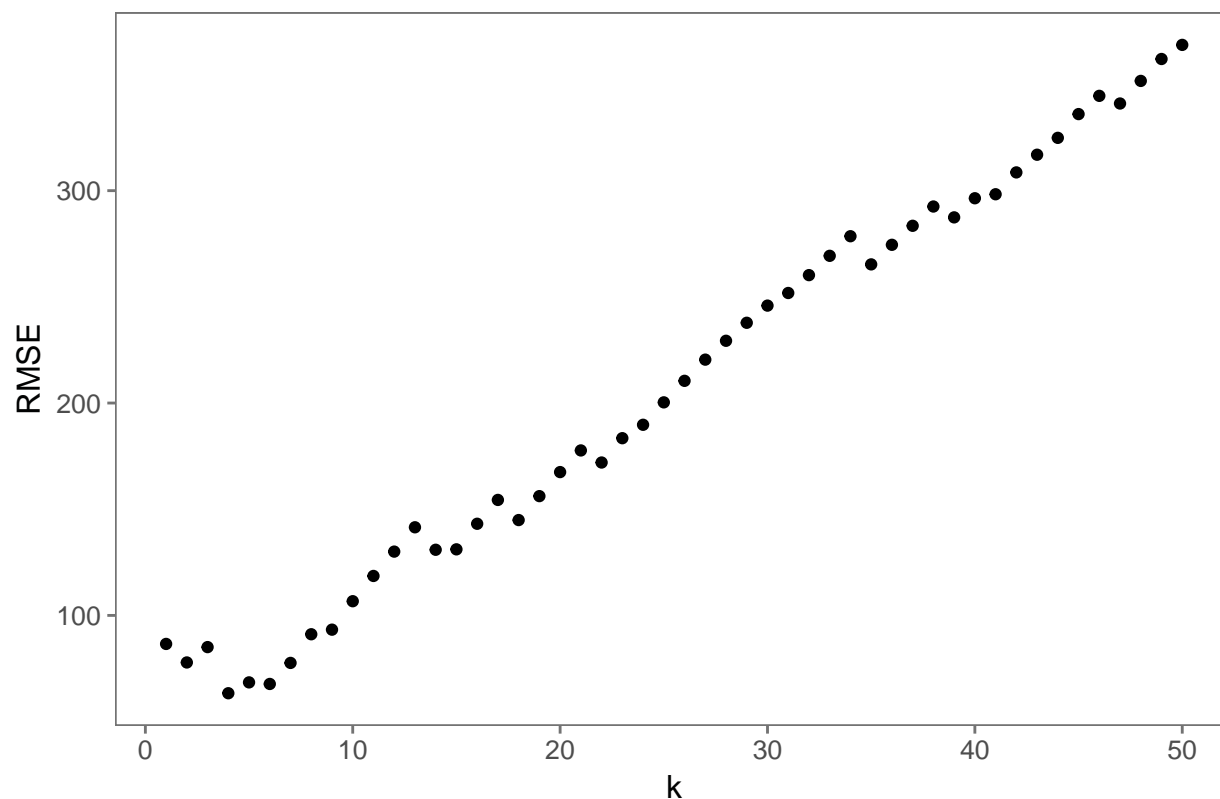
We went ahead and drew a qplot to illustrate the pattern of k from 1 to 50 and its corresponding RMSE.

According to the graph, 4 is the value of k with the least RMSE hence the most accurate in comparison to the others. Above 4, the RMSE gradually increase with the increasing value of k hence overfitting. This can be interpreted to mean that the sale price of each house is closest to that of only the four most similar houses to it, insinuating that there is a high variance in the data as well as high sparsity in the 306 dimensional space of the one-hot encoded data.

KNN regression



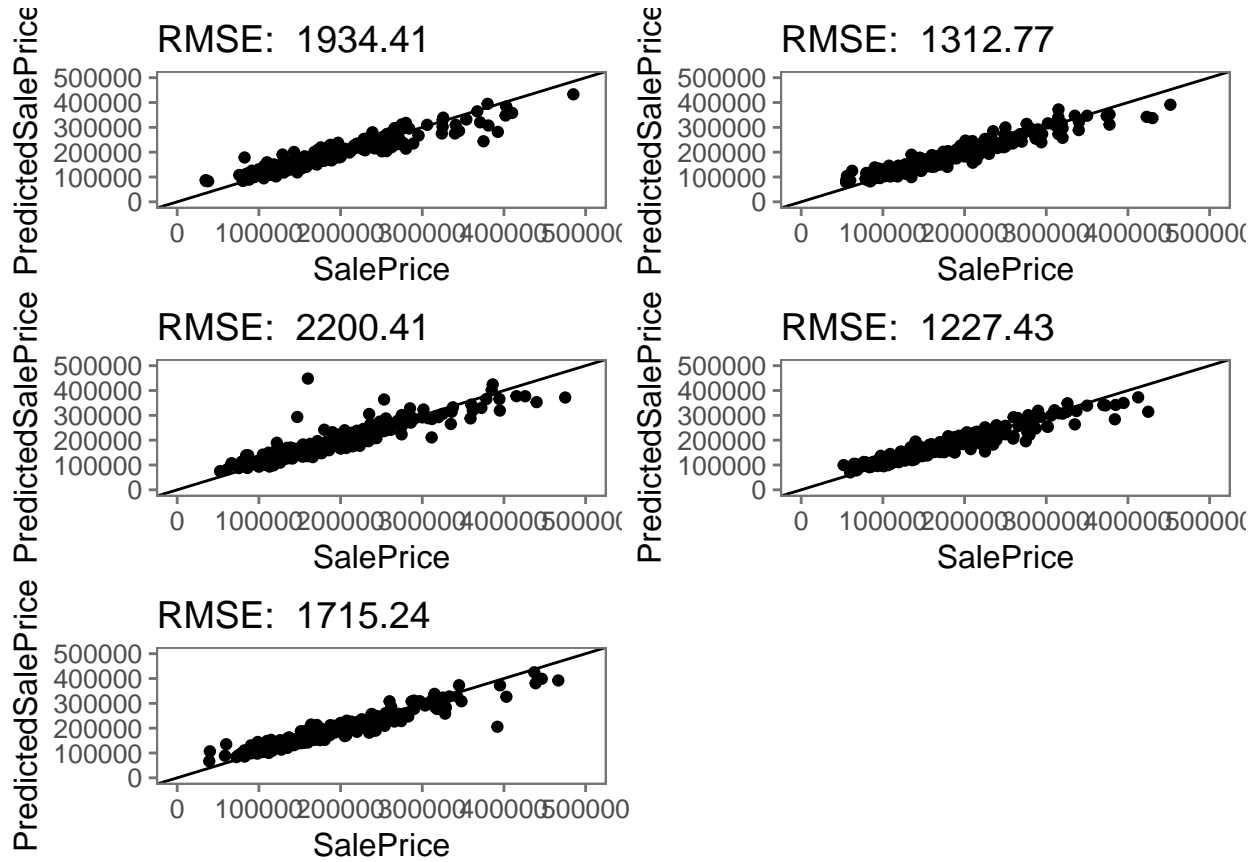
Minimum RMSE at k= 4



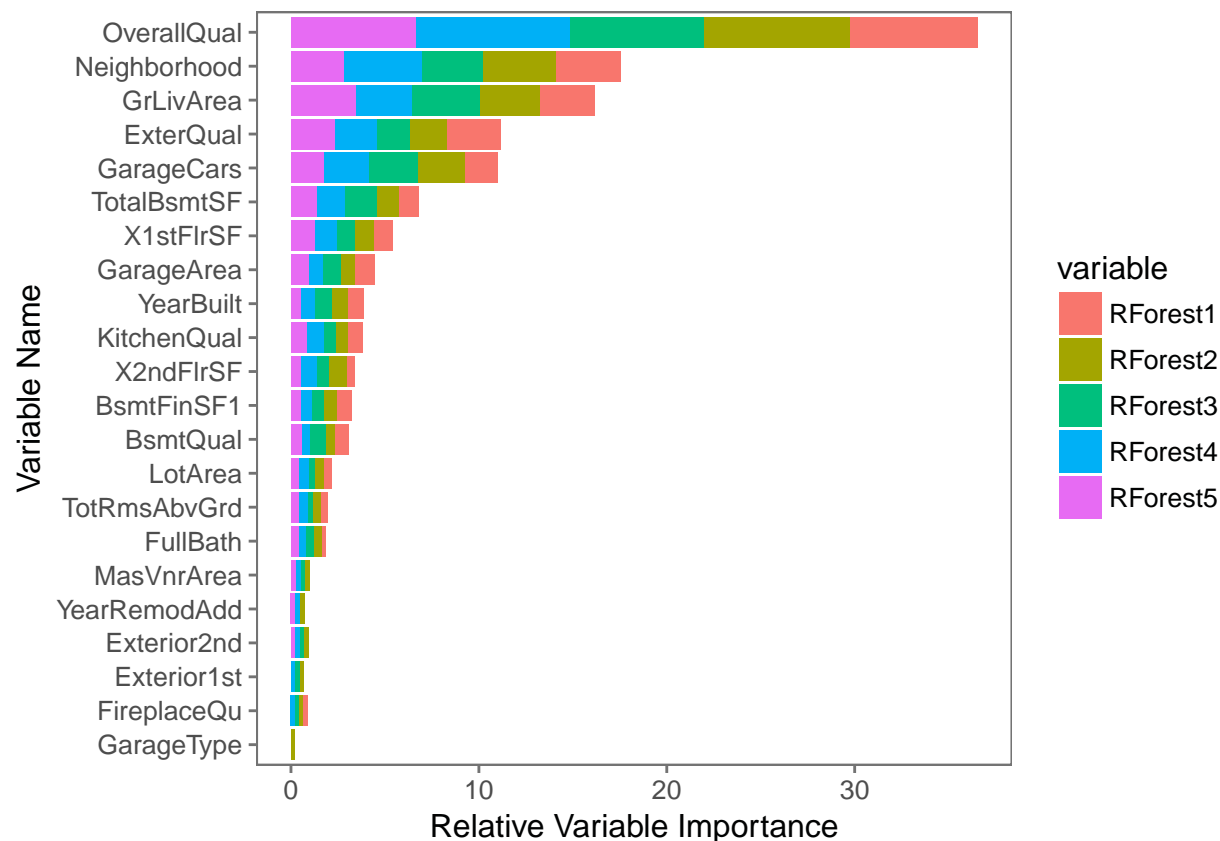
Random Forests

The next method we tried to tackle is the Random Forests, which is an ensemble method based on decision trees. As we have seen, decision trees can be understood easier but they do not predict with the same accuracy as other models. Generally, decision trees have high variance, which might lead to a very different tree for a small change in the training data. Therefore, the random forest allow us to counteract this problem.

In our random forest model, we used 5-fold cross validation to determine how the predictive accuracy could be improved by varying the number of features randomly selected at each node. The data was divided into a training set and test set in order to compare the predicted values with the true values. The figure below shows the target values versus the predicted values. As we can observe, the model is not perfect since all points do not lie along the diagonal line. Indeed, we can notice that the points found in the middle of the x-axis tend to be equally above and below the line. Nevertheless, the modelling for low sale prices or high sale prices was not as accurate. The model tends to overestimate sale prices that are low and underestimate high sale prices.

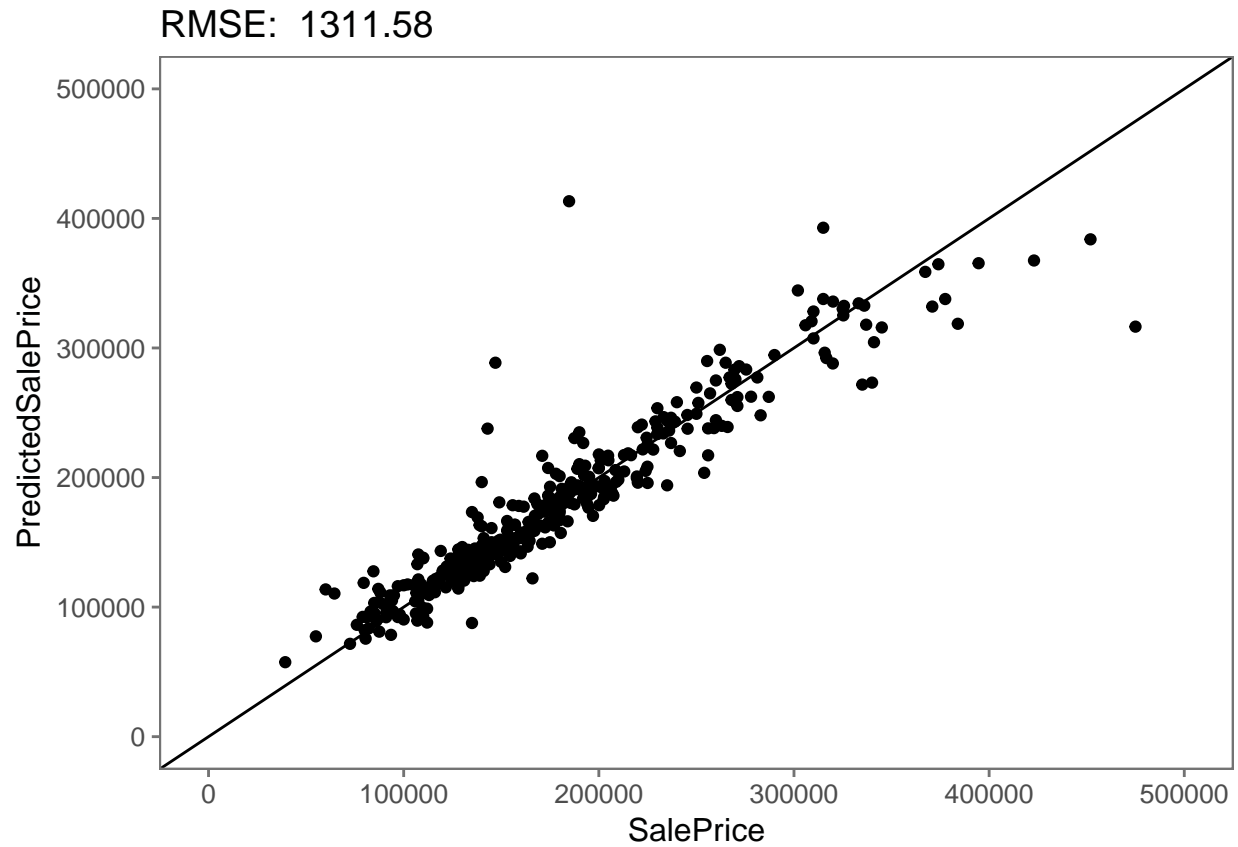


Although random forests are not easily interpretable, they could provide us with the importance of each of the predictors. In our model, we have determined that overall quality was the most important predictor, followed by the house's neighborhood and the square feet of the above ground living area. Nonetheless, the overall quality was clearly the most important one.

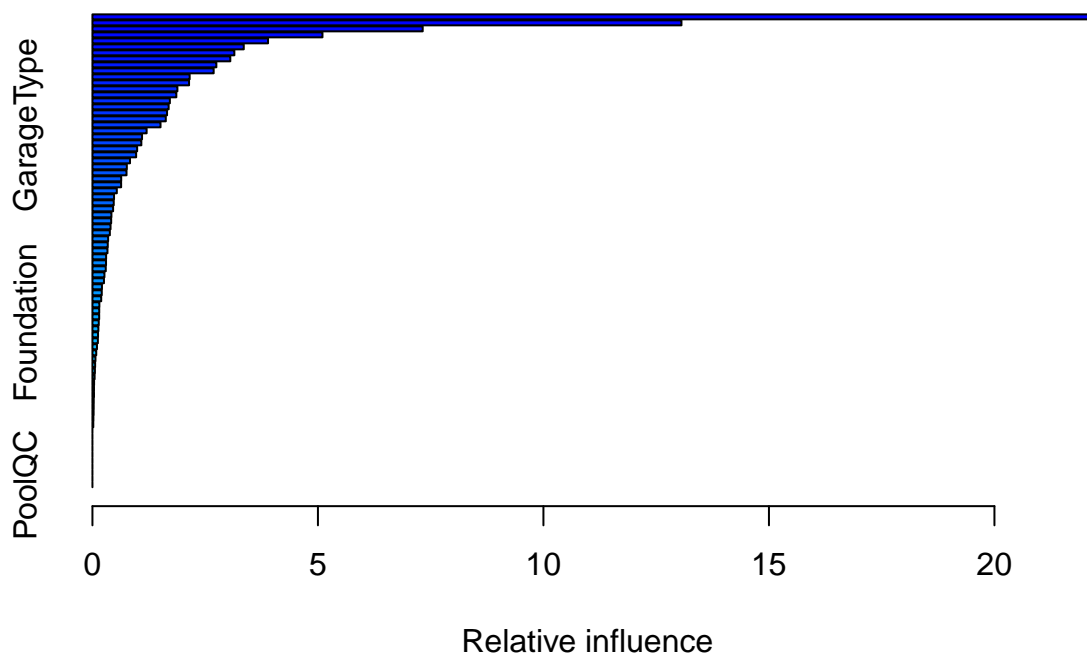


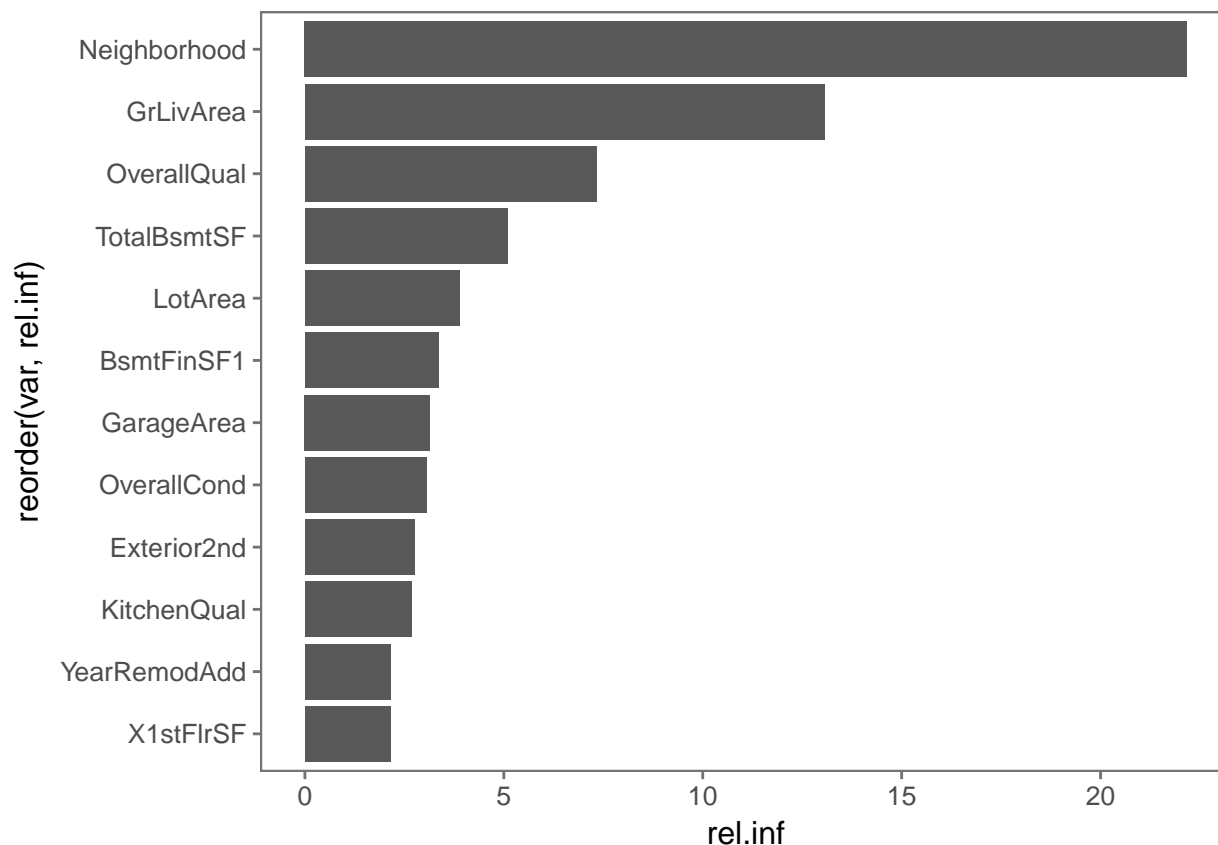
Gradient Boosting Model

Due to the fact that in our random forest model we have overestimated low sale prices and underestimated high sale prices, we can address this bias by applying some boosting techniques (such as the Gradient Boosting Model). As we can observe in the figure below, the GBM model was more accurate in our prediction than the random forests model since it is reduced the bias when overestimating sale prices that are low. However, with this method we did not manage to reduce the bias since it stills underestimate sale prices that are high.



Finally, according to the GBM model, the house's neighborhood is the most important predictor, whereas the the square feet of the above ground living area is the second one, as we can observe in the following plot.





Second Research Question

Then we started looking into answering our second research question: Are there natural clusters of houses that can be derived and explained from within the data and what do those clusters correspond to?

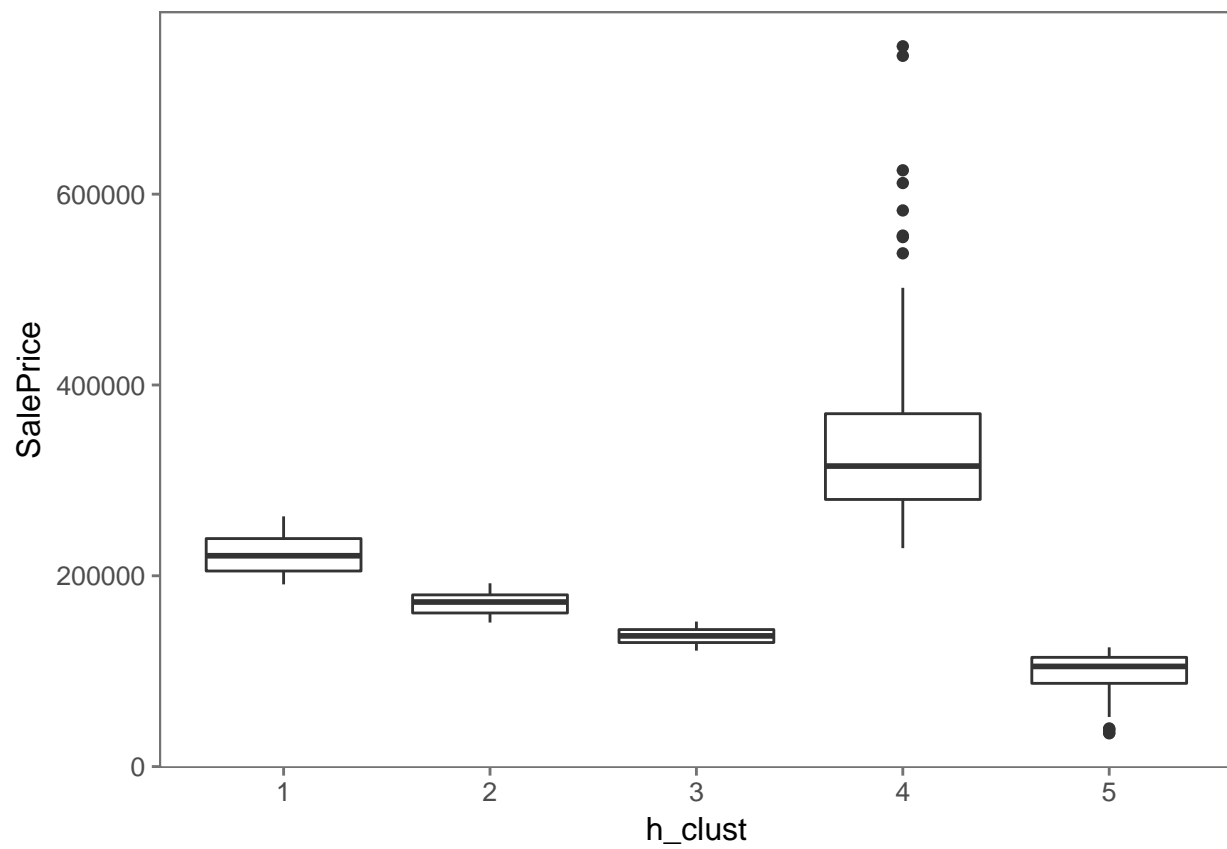
We applied hierarchical clustering in order to classify our data into 5 clusters. This contributes to a part of the answer to our research question “Are there intrinsic groupings/clusters of houses present in the data and what do these clusters represent? The clusters are formed basing on the fact that houses with similar features are within the same range of Sale Price.

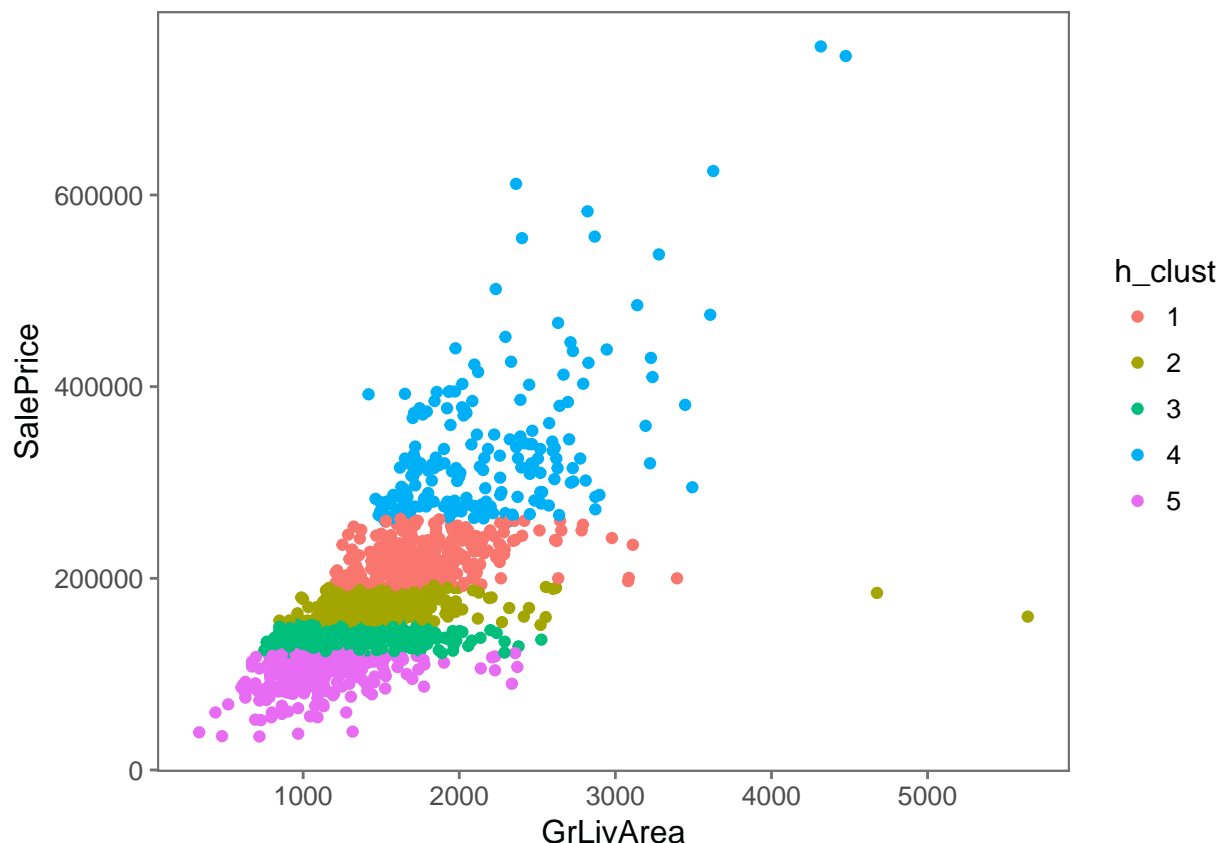
We classified the clusters in 5 groups. The 5 groups represent a range of sale prices in which the dataset fall. The 4th group being the most expensive costing within the range between 300000(3e+05) to 400000(4e+05) with outliers that go beyond 600000(6e+05) while the cheapest group is the 5th group ranging from 100000(1e+05) but less than 200000(2e+05). Moreover we checked the overall quality compared to the price and the 4th group having the highest quality ranging within 7 up to 10 while the 5th group had the lowest quality ranging within 2 up to 6 with a few outliers which have 7.

Moreover, we checked as well the relationship between GrLivArea(Above grade (ground) living area square feet) and the sales price , we found that the 4th group’s GrLiv area ranged between 1500 to 3500 with some outliers reaching to 4000 while the 5th group ranged between 500 to 2500. Interestingly,the second group’s GrLiv area ranges between 1000 till almost 2000 however there are outliers that go beyond 5000.

Cluster Dendrogram





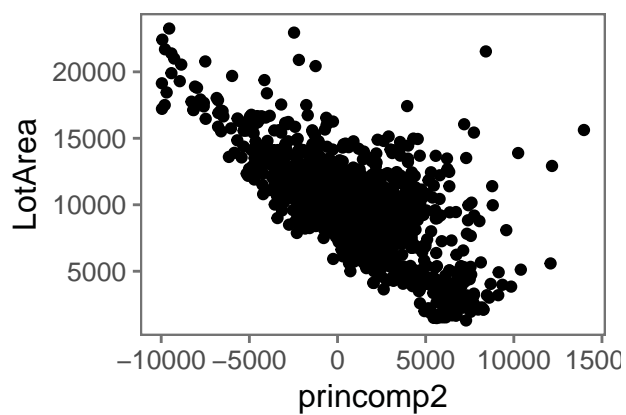
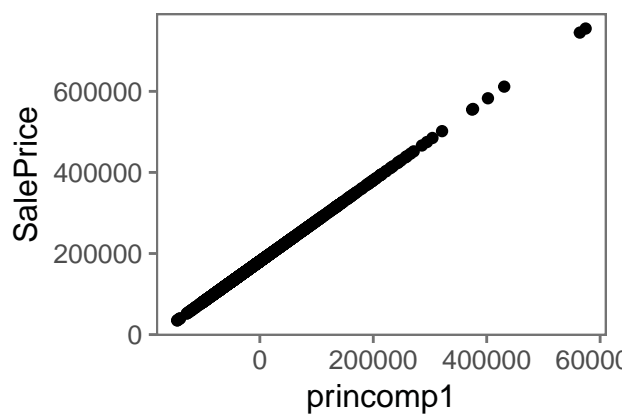
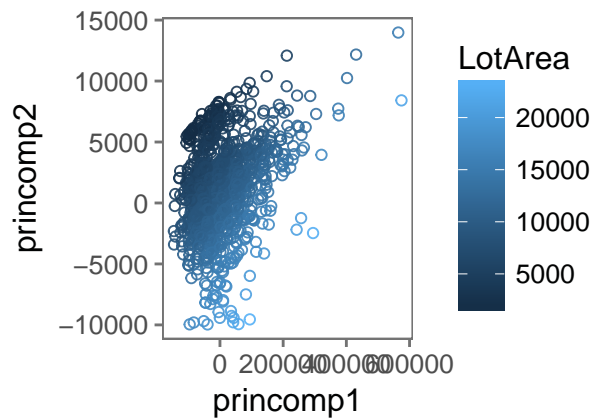
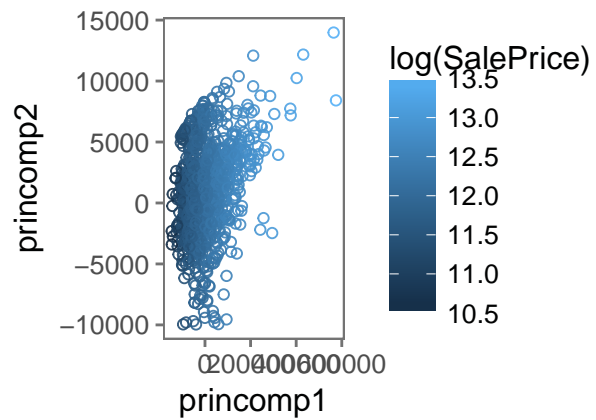


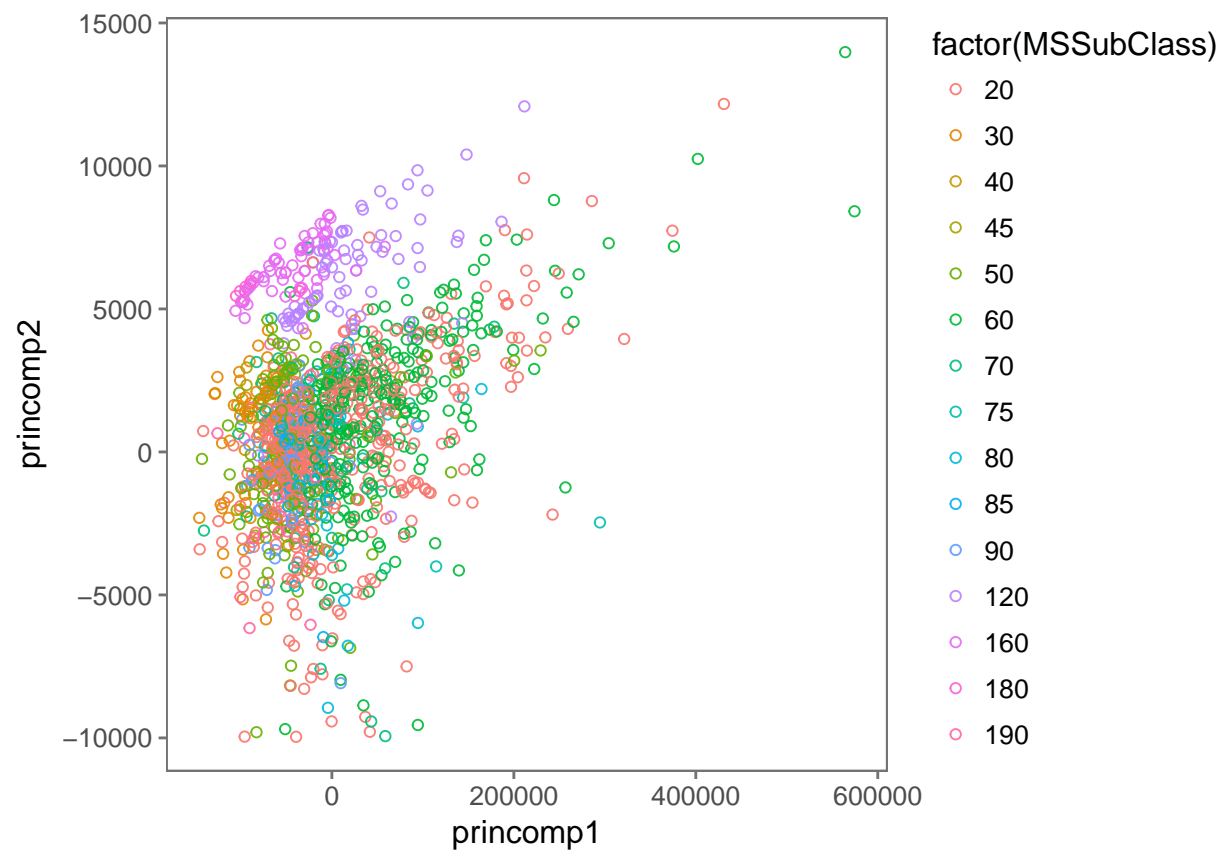
We then decided to do some dimensionality reduction to see if there were any obvious structures in the data.

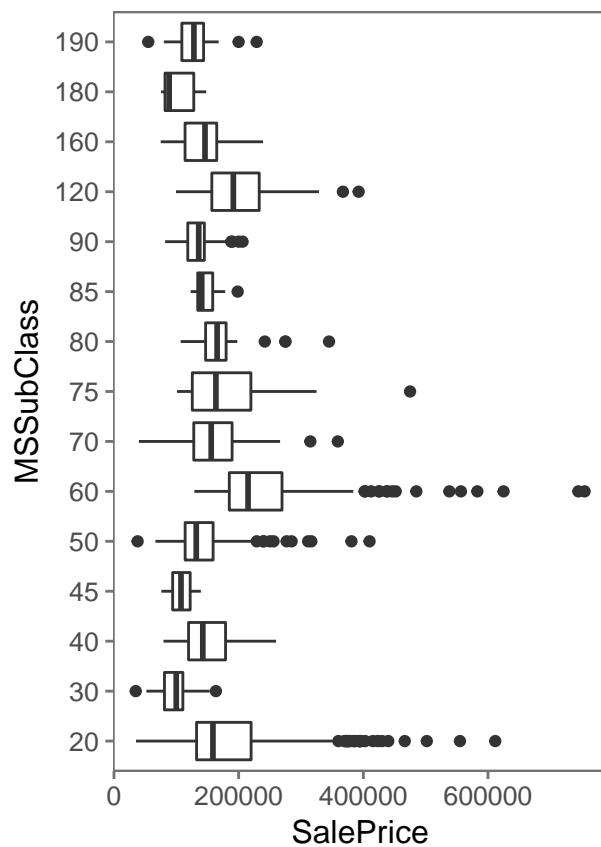
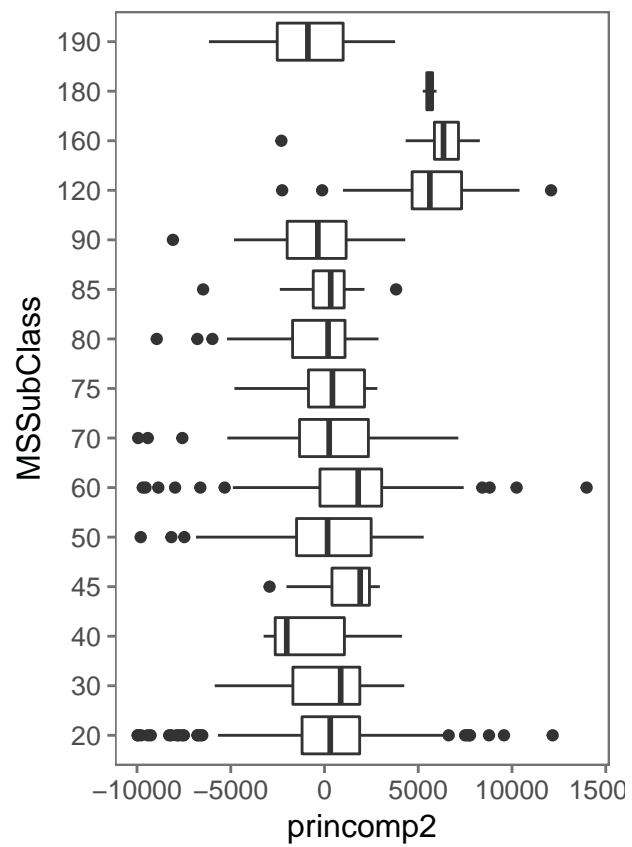
First, we computed PCA on the data with two different algorithms, then plotted them.

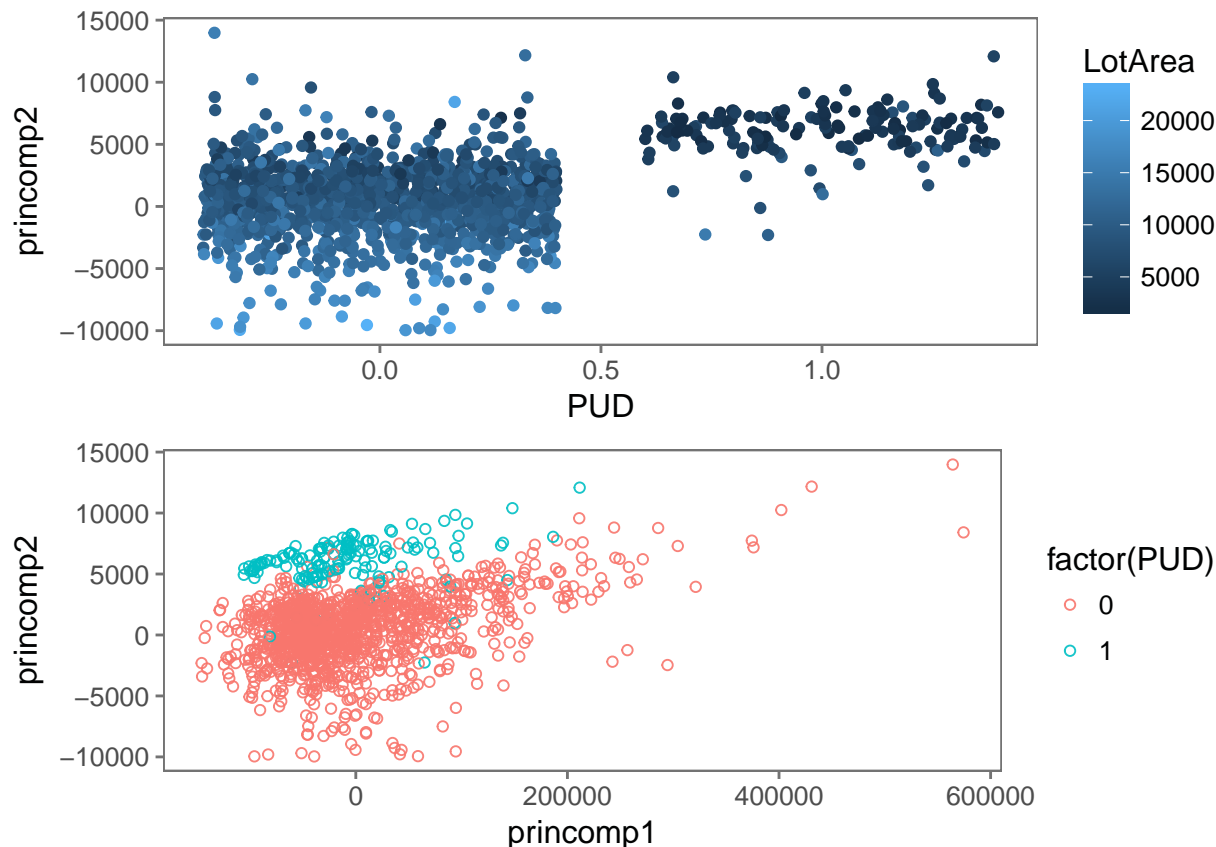
Sale price seems to almost exactly maps to the first principle component, and the second principle component is highly negatively correlated with LotArea. It is therefore obvious that much of the variation within this dataset is explained by these two variables.

Furthermore, MSSubClass (the type of dwelling involved in the sale) also has an interesting relationship with principal component 2. The MSSubClasses 120, 160, and 180 form a cluster in the higher values of the second principal component. MSSubClasses 120, 160, and 180 are the 3 types of dwellings that are planned unit developments - communities of homes that are operated by a homeowners association to provide amenities like parks, playgrounds, pools, tennis and basketball courts, hiking trails, private gated common land and street lights or the like. It makes sense that these homes form a cluster, because homes in planned communities are often regulated to be extremely similar.





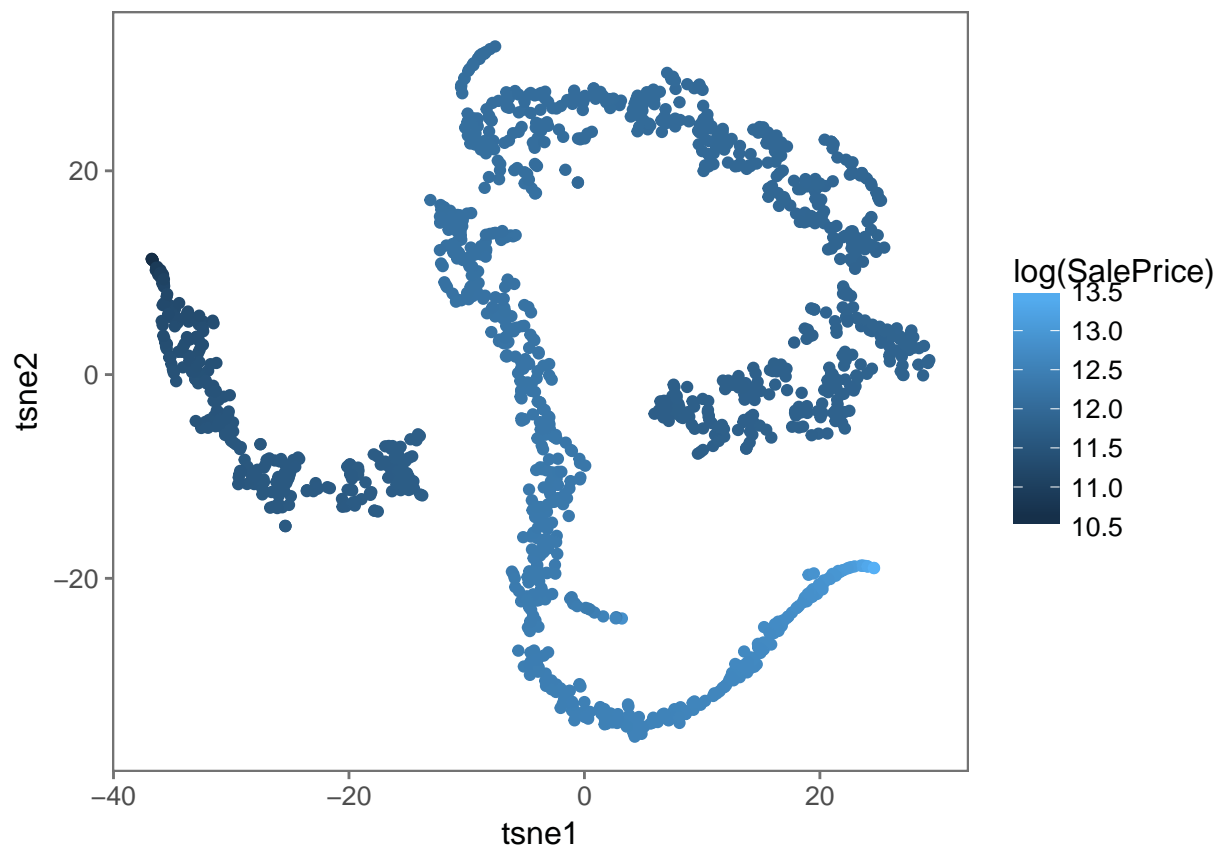


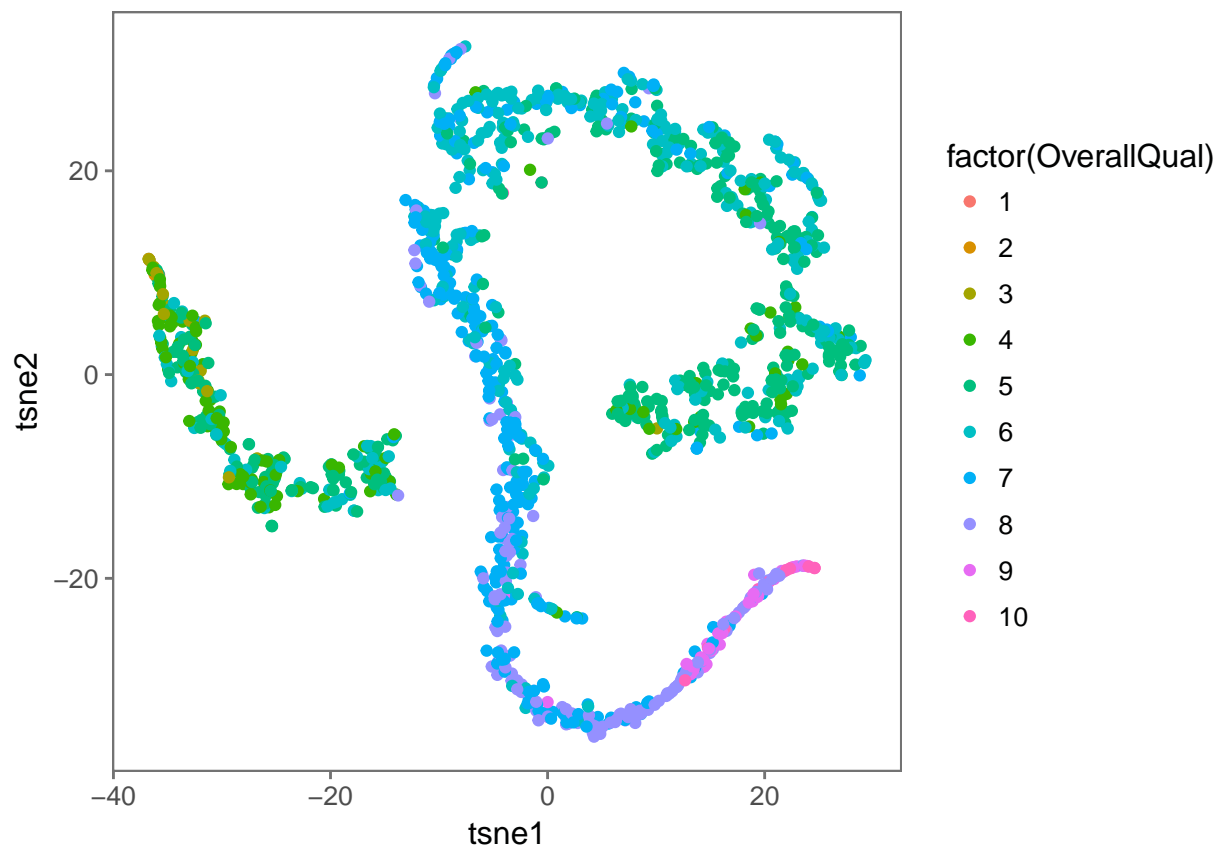


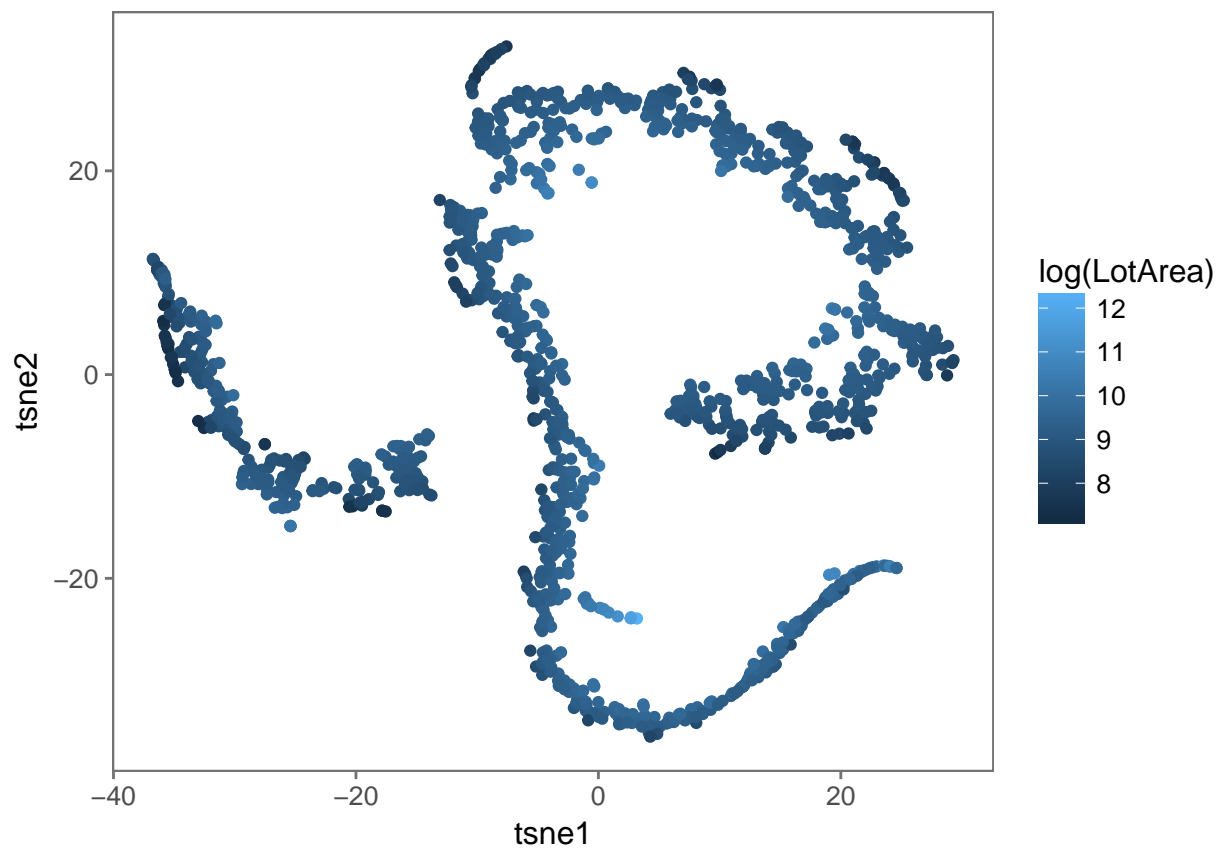
Then we decided to compare dimensionality reduction of a t-SNE plot, which stands for t-distributed Stochastic Neighbor Embedding. It is a dimensionality reduction algorithm that embeds high dimensional data into a 2 dimensional space. There is a nice R vignette about the package, a blog post about its use in R, and an interactive visualization characterizing its strengths and weaknesses.

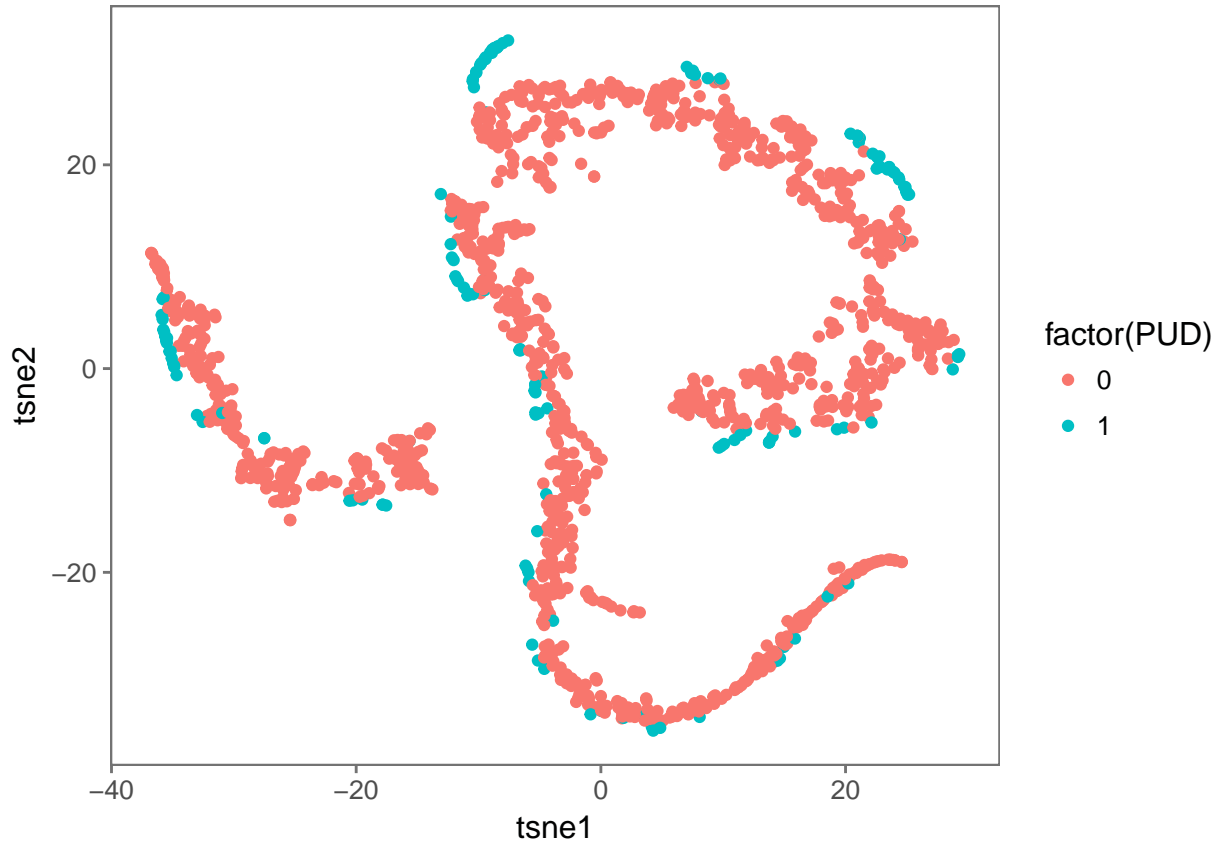
The t-SNE confirms many of the same things as the principal component analysis, namely that the majority of the variance in the dataset can be explained by the sale price (as well as variables that are correlated to the sale price like the overall quality). Also, when you plot variables that are uncorrelated with the SalePrice but correlated with the second principal component (like lot area and being a PUD) you see groupings on the outside of the main structure, and the occasional small cluster. One cluster (seen in the t-SNE plot colored by lot area) seems to point to the similarity of more rural houses that have much higher lot area. Coloring the t-SNE plot by the PUD houses also shows some clear clusters of houses likely in the same housing development.

```
## Joining, by = c("Id", "MSSubClass", "LotFrontage", "LotArea", "OverallQual", "OverallCond", "YearBui
```







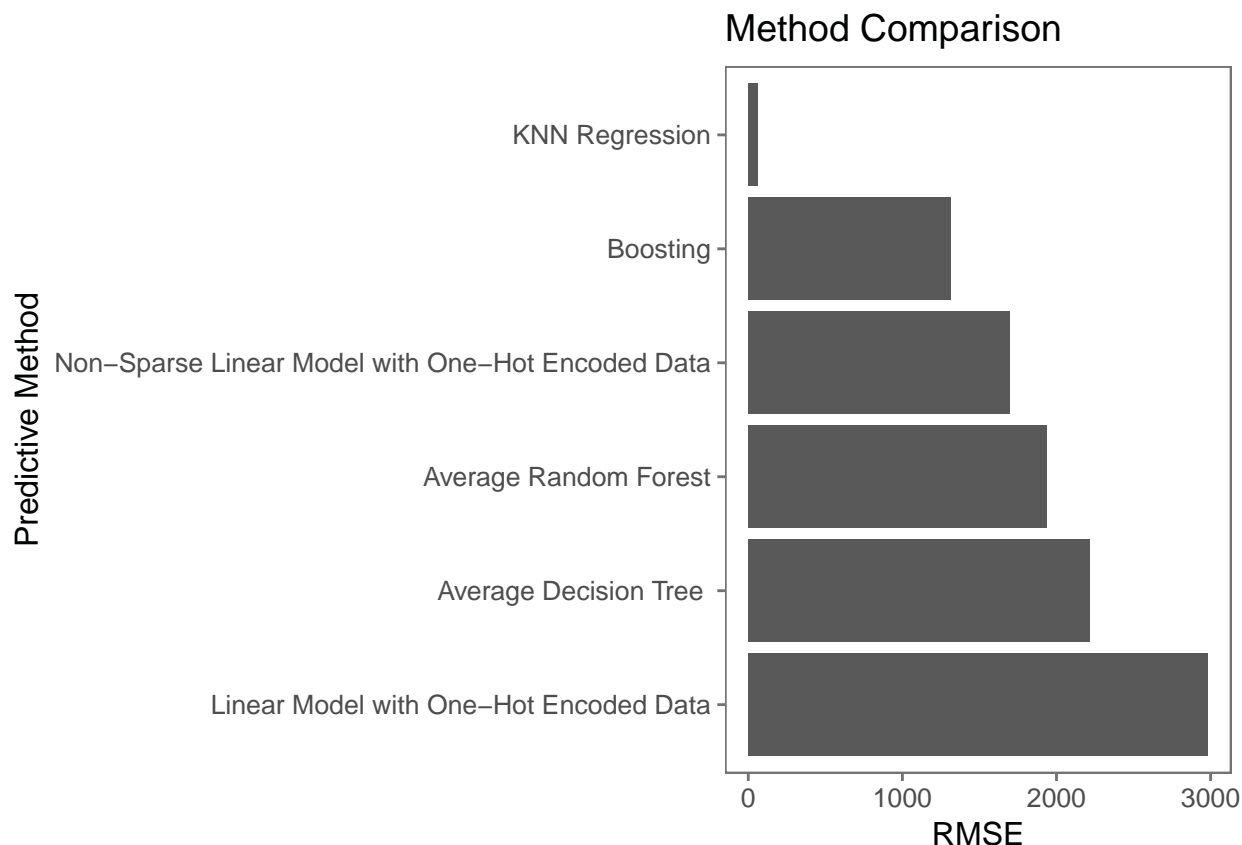
Overall Conclusion:

Overall, the answers to our research questions were interesting and diverse.

Research Question 1:

For the first research question, there are many factors that are important to predicting the sales price, but it seems like the most important (in no particular order) are the overall house quality, above grade living area square feet, neighborhood, external quality, size of garage in car capacity, and kitchen quality. These variables were determined to be important by many of the predictive algorithms we examined as well as by the exploratory data analysis and correlation analysis at the beginning of the project.

From a comparison of the RMSEs of all the different algorithms, it's obvious that the knn regression algorithm is the best fit for this data. This is likely due to the high variability in the characteristics of houses, which makes it difficult to come up with overall rules that apply to all situations like regression trees do with variable splitting and linear models do with coefficients.



Research Question 2:

From our analysis of reducing the dimensionality reduction of our data, it seems as though there are two main groupings within our data. The first is the sale price of the home. The sale price of a home is highly correlated with many other variables in the dataset and knowing the sale price can give you a good idea about other details of the home. Our analysis of hierarchical clustering isolated five clusters that map almost exactly to subsets of the sale price. The second main grouping is the amount of lot area/PUD status of a home. These groupings likely are a proxy for measuring how suburban/rural a home is, because suburban homes are often PUDs that have little individual lot area, but share land within the subdivision, and rural homes often have much more land than their city counterparts.

The price and suburban-ness of a home make sense as good metrics for explaining the variation between homes because an expensive loft in the city is quite different than an expensive suburban or rural home, as is an inexpensive suburban/rural home from an inexpensive urban home.

In short, our two research questions uncovered some interesting results from our data and we hope to post our findings on Kaggle in order to get feedback from and share our findings with the wider data science community.