Data Visualization and Analysis

Semester Project

Amrit Kaur
Pratik Agarwal
Mayuri Mendke
Nofel Mahmood

Dataset

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Motivation

The Dataset has 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. The problem was to predict sale prices of the houses based on the variables. The problem was interesting to us because we had studied linear regression and other models to solve problems like this.

Approach

As there were a lot of variables we tried to first find out the ones which were influencing the house price on a higher level and then used them to build our linear model to predict house sale prices. After that we applied time series to forecast house prices for the next 10 years. In the end we applied clustering to group houses with respect to sale price (low, mid, high) and use inference tree to show sale prices with respect to years in which the particular house was built.

Exploration / Visualizations

Install Pacman

```
library(pacman)
```

Load all required packages

```
p_load(tidyverse, stringr,lubridate, ggplot2, tseries, forecast, scales, party)

house_training_data <- read.csv("./DataSet/train.csv")
house_test_data <- read.csv("./DataSet/test.csv")
```

Get the idea of Minimum and Maximum Price of the house along with mean and others.

```
summary(house_training_data$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

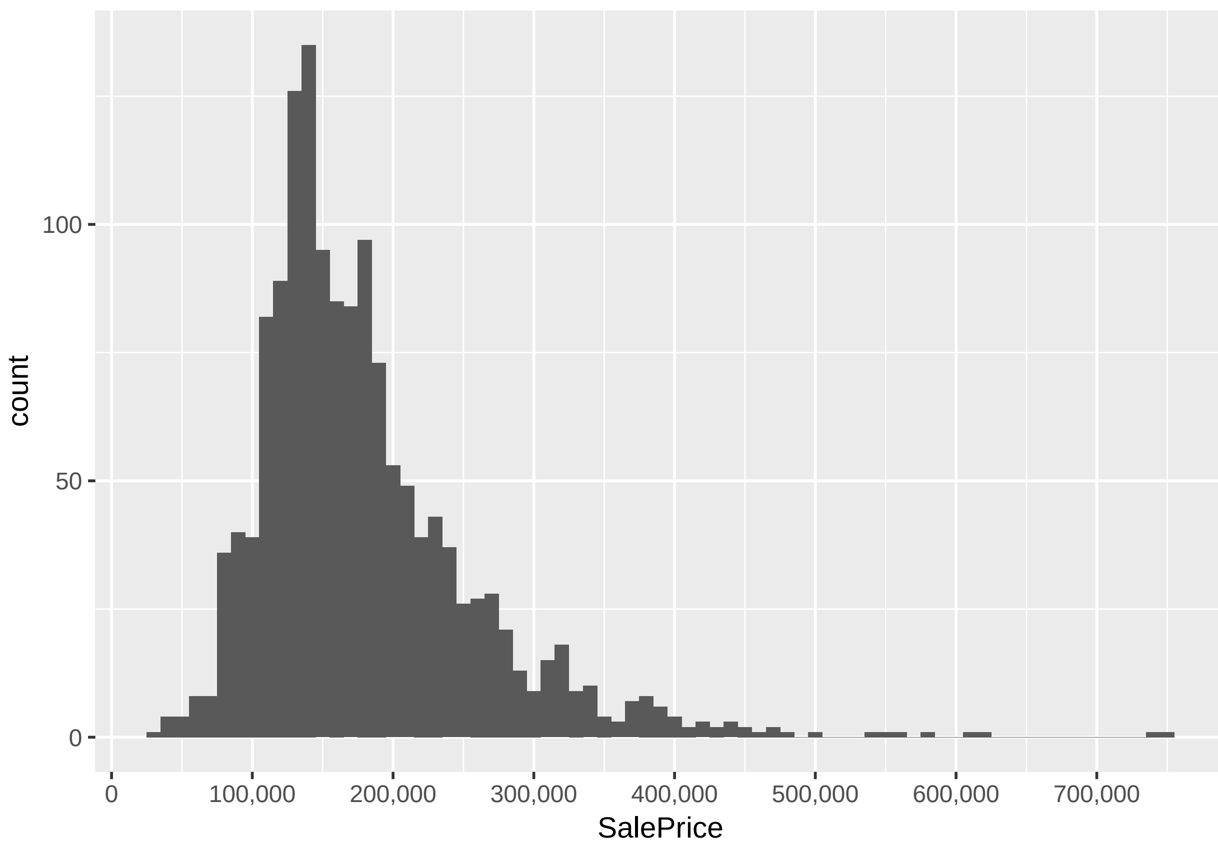
Add saleprice column to the test data. And assigned it to a new variable. Combine both the training and test data. It will be easier for analysis. From now on we will work on this dataset.

```
house_test_data.SalePrice <-
  data.frame(SalePrice = rep(NA, nrow(house_test_data)), house_test_data[,])

house_test_data.SalePrice <-
  data.frame(SalePrice = rep(NA, nrow(house_test_data)), house_test_data[,])
house_combined <- rbind(house_training_data, house_test_data.SalePrice)

dim(house_combined)  #Dimention of the combined dataset.
```

```
## [1] 2919   81
```

With this plot we can say: Few people can afford very expensive houses. Majority of people bought houses in the range 1,00,000 to 2,50,000.

```
training_data <- house_training_data[!is.na(house_combined$SalePrice),]
training_data %>% ggplot(aes(x=SalePrice)) +
                  geom_histogram(binwidth = 10000) +
                  scale_x_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```



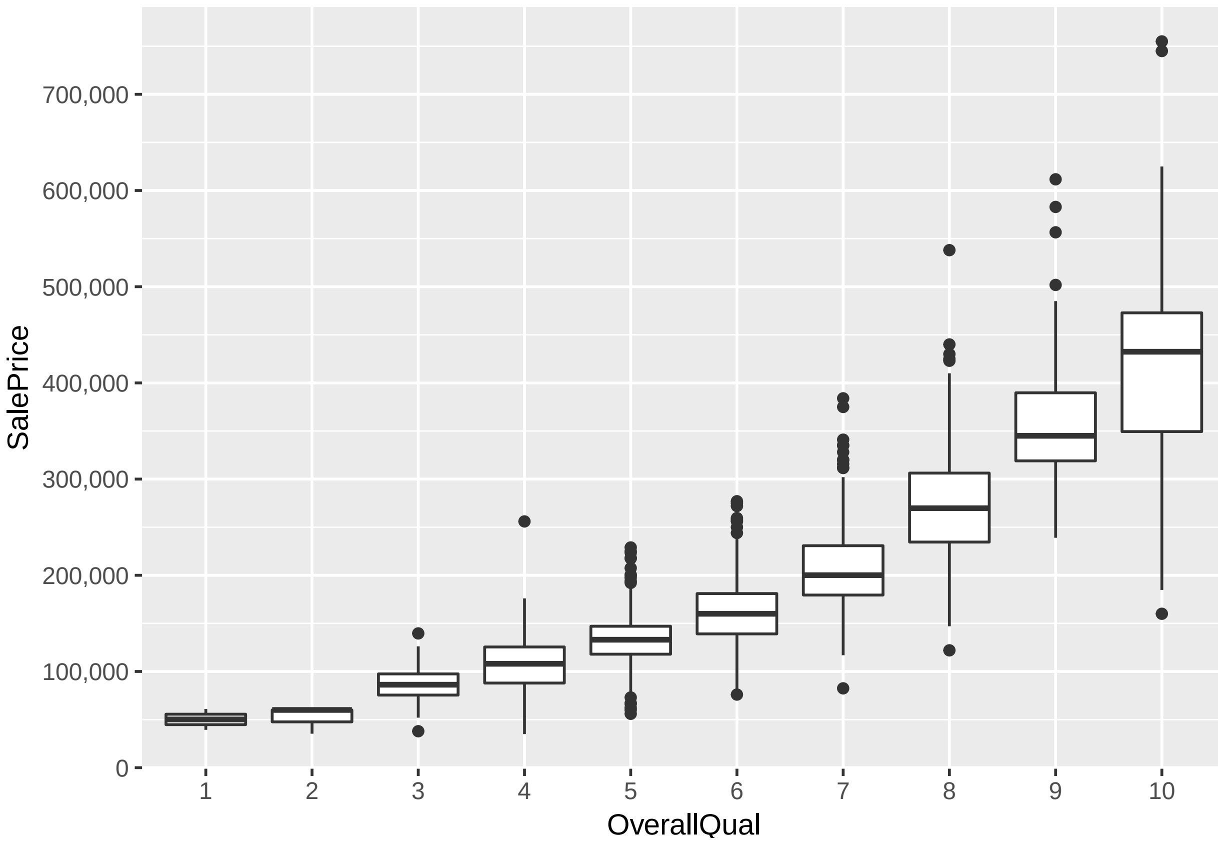
Now we have to find which attributes are more significant for SalePrice.

```
#We don't need ID. So drop ID column from house_combined
house_training_data$Id <- NULL

#Here we have selected only those variables which has type numeric.
#Now we can check there correlation with SalePrice.
numeric.type.variables <- which(sapply(house_training_data, is.numeric))
numeric.type.name.variables <- names(numeric.type.variables)
```

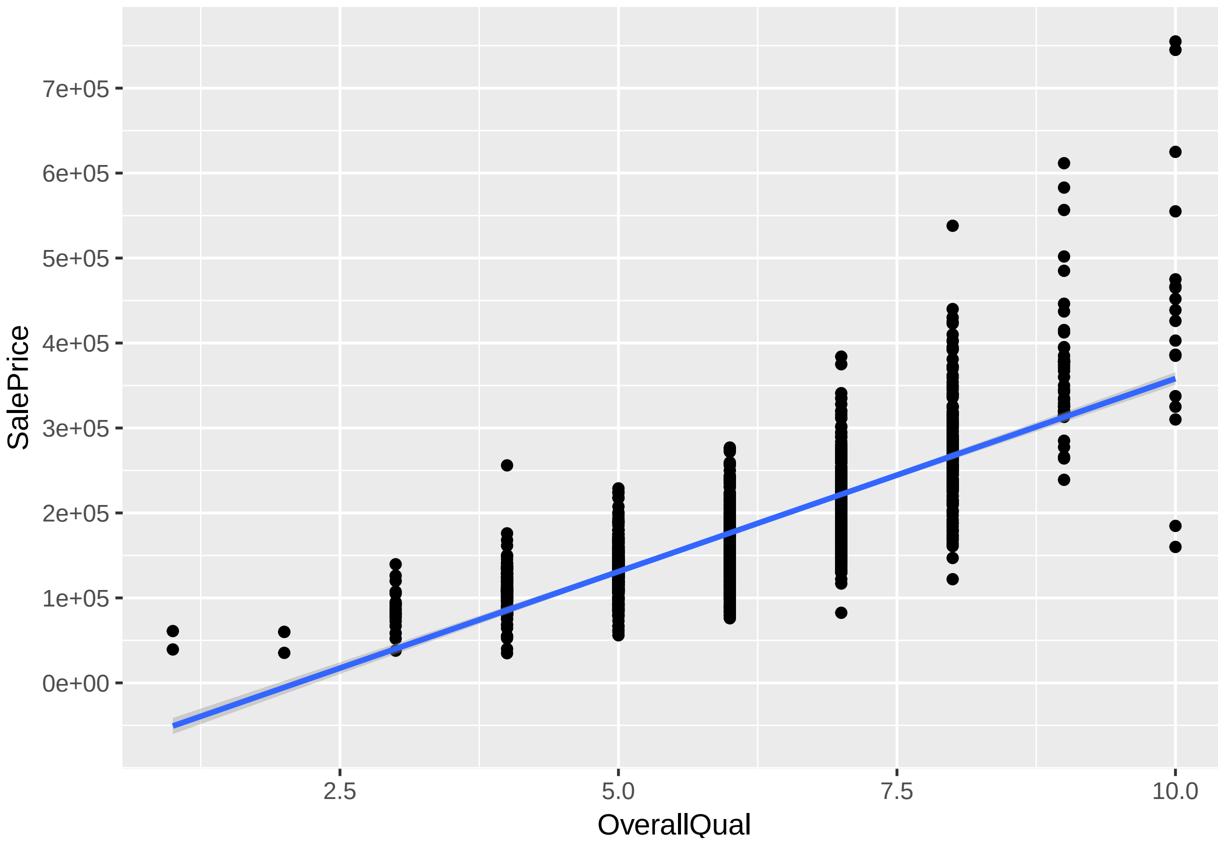
```r
cor.numeric.variables <- cor(house_training_data[, numeric.type.variables],
                             use="pairwise.complete.obs")
#Lot of NA's .
#so we use="pairwise.complete.obs".

#sort the correlation with saleprice in decreasing order.
#So we will get the highly correlated variable at the top.
cor_sorted <- as.matrix(sort(cor.numeric.variables[,'SalePrice'], decreasing = TRUE))
colnames(cor_sorted)<- c("values")

#Select only high correlation
CorHigh <- names(which(apply(cor_sorted, 1, function(x) abs(x)>0.5)))
#So we got "OverallQual" as the highly significant variable for Saleprice and after that
#we "GrLivArea" and so on..

model_OverallQual<-lm(SalePrice~OverallQual, data = house_training_data)
summary(model_OverallQual)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual, data = house_training_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -198152  -29409   -1845   21463  396848
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -96206.1     5756.4  -16.71   <2e-16 ***
## OverallQual  45435.8      920.4   49.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48620 on 1458 degrees of freedom
## Multiple R-squared:  0.6257, Adjusted R-squared:  0.6254
## F-statistic:  2437 on 1 and 1458 DF,  p-value: < 2.2e-16
```

```r
ggplot(house_training_data[!is.na(house_training_data$SalePrice),],
       aes(x= factor(OverallQual), y = SalePrice)) +
       geom_boxplot() + labs(x = "OverallQual", y = "SalePrice") +
       scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```




```
#We can Clearly see that increase in the overall quality of the house has increased the
#saleprice of the house.
house_training_data %>% ggplot(aes(x=OverallQual, y= SalePrice)) +
                        geom_point() +
                        geom_smooth(method = "lm") +
                        scale_y_continuous(breaks = seq(0,800000, by=100000))
```

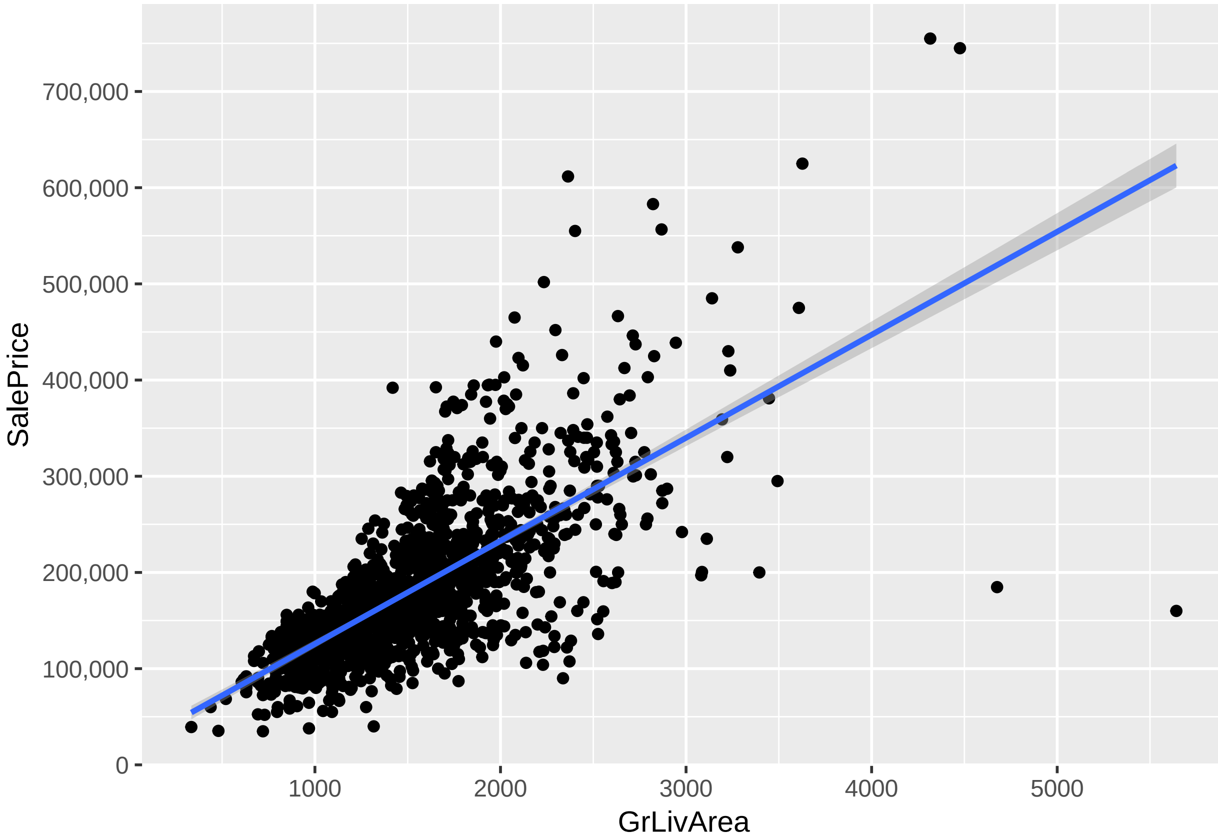


Our models

```
model_GrLiveArea<-lm(SalePrice~GrLivArea, data = house_training_data)
summary(model_GrLiveArea)
```

```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea, data = house_training_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -462999  -29800   -1124   21957  339832
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18569.026   4480.755   4.144 3.61e-05 ***
## GrLivArea     107.130      2.794  38.348  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56070 on 1458 degrees of freedom
## Multiple R-squared:  0.5021, Adjusted R-squared:  0.5018
## F-statistic:  1471 on 1 and 1458 DF,  p-value: < 2.2e-16
```

```r
ggplot(house_training_data[!is.na(house_training_data$SalePrice),],
       aes(x= GrLivArea, y = SalePrice)) + geom_point() +
       geom_smooth(method = "lm") +  labs(x = "GrLivArea", y = "SalePrice") +
       scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```




```r
#Next highly correlated variable was "GrLivArea" i.e ground living area square feet.
#Which also makes sense as the house with
#bigger living area will have high sale price.
#The two dots at the bottom right seems to be the outliers.

factor.type.variables.names<- which(sapply(house_training_data, is.factor))%>% names()
model_Street_Neighborhood <- lm(SalePrice ~ Street+Neighborhood+GarageCond+
                                  KitchenQual+MiscFeature,
                                data = house_training_data)
summary(model_Street_Neighborhood)
```

```
##
## Call:
## lm(formula = SalePrice ~ Street + Neighborhood + GarageCond +
##     KitchenQual + MiscFeature, data = house_training_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47070 -17561      0  14025  94121
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           253266      65834   3.847 0.000580 ***
## StreetPave           -134950      38966  -3.463 0.001628 **
## NeighborhoodClearCr   128389      34430   3.729 0.000799 ***
## NeighborhoodCollgCr   114820      31750   3.616 0.001083 **
## NeighborhoodCrawfor   156389      41890   3.733 0.000790 ***
## NeighborhoodEdwards    17132      29905   0.573 0.571006
## NeighborhoodGilbert    83296      28026   2.972 0.005781 **
## NeighborhoodIDOTRR   -156569      43968  -3.561 0.001255 **
## NeighborhoodMitchel    57250      25774   2.221 0.034028 *
## NeighborhoodNAmes      40727      21943   1.856 0.073294 .
## NeighborhoodNWAmes     88722      26059   3.405 0.001900 **
## NeighborhoodOldTown    67879      25612   2.650 0.012714 *
## NeighborhoodSawyer     34149      26383   1.294 0.205422
## NeighborhoodSawyerW    78889      41890   1.883 0.069397 .
## NeighborhoodTimber        NA         NA      NA       NA
## GarageCondFa           -9278      45071  -0.206 0.838298
## GarageCondTA            7227      35223   0.205 0.838814
## KitchenQualGd          23879      32403   0.737 0.466887
## KitchenQualTA           7768      29998   0.259 0.797427
## MiscFeatureOthr       -41039      42463  -0.966 0.341535
## MiscFeatureShed       -39312      26147  -1.503 0.143174
## MiscFeatureTenC        11855      48023   0.247 0.806694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33750 on 30 degrees of freedom
##   (1409 observations deleted due to missingness)
## Multiple R-squared:  0.7424, Adjusted R-squared:  0.5707
## F-statistic: 4.323 on 20 and 30 DF,  p-value: 0.000161
```

```r
#Here, we can see that street and Neighbourhood are significant variables effecting the
#SalesPrice of an house. The dummy Variables StreetPave,NeighborhoodCollgCr,
#NeighborhoodCrawfor are most significant.
#The model is very good because it has a high R value.

lm(SalePrice~Street+Neighborhood+OverallQual+
             GrLivArea+Condition1+Condition2+
             SaleCondition+SaleType+Heating,
             data = house_training_data)
```

```
##
## Call:
## lm(formula = SalePrice ~ Street + Neighborhood + OverallQual +
##     GrLivArea + Condition1 + Condition2 + SaleCondition + SaleType +
##     Heating, data = house_training_data)
##
## Coefficients:
##         (Intercept)           StreetPave   NeighborhoodBlueste
##           -25185.70             -1442.17             -18807.11
##   NeighborhoodBrDale   NeighborhoodBrkSide   NeighborhoodClearCr
##           -31895.47             -2638.45              34956.05
##  NeighborhoodCollgCr   NeighborhoodCrawfor   NeighborhoodEdwards
##            16271.44             22658.71              -6722.27
##  NeighborhoodGilbert    NeighborhoodIDOTRR   NeighborhoodMeadowV
##             2392.89            -16442.81             -11210.14
```

```
##  NeighborhoodMitchel     NeighborhoodNAmes   NeighborhoodNPkVill
##             10701.79               6581.36              -7801.88
##   NeighborhoodNWAmes   NeighborhoodNoRidge   NeighborhoodNridgHt
##              5113.32              73038.91              69746.00
##  NeighborhoodOldTown     NeighborhoodSWISU    NeighborhoodSawyer
##            -17924.74             -25587.05              11538.68
##  NeighborhoodSawyerW   NeighborhoodSomerst   NeighborhoodStoneBr
##             11594.88              15128.22              69858.52
##   NeighborhoodTimber   NeighborhoodVeenker           OverallQual
##             33759.26              56848.92              20264.92
##            GrLivArea       Condition1Feedr        Condition1Norm
##                57.09                 -4.58              15426.90
##       Condition1PosA        Condition1PosN        Condition1RRAe
##             22940.80              19254.05              -3305.96
##       Condition1RRAn        Condition1RRNe        Condition1RRNn
##             19334.02               5007.90              34366.33
##      Condition2Feedr        Condition2Norm        Condition2PosA
##            -45285.02             -17095.05              20926.20
##       Condition2PosN        Condition2RRAe        Condition2RRAn
##           -151875.73             -33359.28             -26365.46
##       Condition2RRNn  SaleConditionAdjLand  SaleConditionAlloca
##             16523.87               -499.98               8316.26
##  SaleConditionFamily   SaleConditionNormal  SaleConditionPartial
##             -2532.47               8890.56              12428.04
##          SaleTypeCWD           SaleTypeCon         SaleTypeConLD
##             16679.42              37215.44              11981.13
##        SaleTypeConLI         SaleTypeConLw           SaleTypeNew
##             17032.06              -3539.25              33986.79
##          SaleTypeOth            SaleTypeWD           HeatingGasA
##             23843.50               8076.71             -29411.44
##          HeatingGasW           HeatingGrav           HeatingOthW
##            -32017.43             -39096.52            -112902.99
##          HeatingWall
##            -37105.36
```

```
#Here, we saw Condition1, Condition2, Saletype, SaleCondition and Heating are not
#that significant when grouped with Street and Neighbourhood. So we will not take
#this model. Also OverallQual and GrLiveArea when added to Street and Neighborhood
#are seen as significant for SalePrice. So we will group these 4 together and
#consider them to train our model for further prediction.

lm(SalePrice~Street+Neighborhood+OverallQual+GrLivArea+
             Exterior1st+Exterior2nd+GarageCars+GarageArea+
             BsmtQual+TotalBsmtSF+BsmtCond+FullBath+
             BsmtFinType1+BsmtFinType2, data = house_training_data)
```

```
##
## Call:
## lm(formula = SalePrice ~ Street + Neighborhood + OverallQual +
##     GrLivArea + Exterior1st + Exterior2nd + GarageCars + GarageArea +
##     BsmtQual + TotalBsmtSF + BsmtCond + FullBath + BsmtFinType1 +
##     BsmtFinType2, data = house_training_data)
##
## Coefficients:
##         (Intercept)           StreetPave  NeighborhoodBlueste
```

```
##            25269.483             2695.116            -16071.952
##   NeighborhoodBrDale  NeighborhoodBrkSide  NeighborhoodClearCr
##           -25037.480             5124.356             29197.199
##  NeighborhoodCollgCr  NeighborhoodCrawfor  NeighborhoodEdwards
##            15347.583            32174.765            -7316.348
##  NeighborhoodGilbert   NeighborhoodIDOTRR  NeighborhoodMeadowV
##            14686.967            -8754.840            -23037.626
##  NeighborhoodMitchel    NeighborhoodNAmes  NeighborhoodNPkVill
##            -7177.406              747.122            -11888.258
##   NeighborhoodNWAmes  NeighborhoodNoRidge  NeighborhoodNridgHt
##              622.942            66816.939             50481.633
##  NeighborhoodOldTown    NeighborhoodSWISU   NeighborhoodSawyer
##           -11626.256            -8937.911              3856.892
##  NeighborhoodSawyerW  NeighborhoodSomerst  NeighborhoodStoneBr
##            10836.533            22206.571             59187.250
##   NeighborhoodTimber  NeighborhoodVeenker          OverallQual
##            23744.557            42513.610             13637.533
##            GrLivArea   Exterior1stBrkComm   Exterior1stBrkFace
##               47.861           -44489.698             13702.967
##    Exterior1stCBlock   Exterior1stCemntBd   Exterior1stHdBoard
##            -2576.129            46700.060             -8613.529
##   Exterior1stImStucc   Exterior1stMetalSd   Exterior1stPlywood
##           -65232.085              281.066             -4418.558
##     Exterior1stStone    Exterior1stStucco   Exterior1stVinylSd
##             8982.767             3391.264            -15438.501
##   Exterior1stWd Sdng   Exterior1stWdShing   Exterior2ndAsphShn
##            -6738.139             3287.062             16106.468
##   Exterior2ndBrk Cmn   Exterior2ndBrkFace    Exterior2ndCBlock
##             3292.582            10590.079                   NA
##   Exterior2ndCmentBd   Exterior2ndHdBoard   Exterior2ndImStucc
##           -25370.492            12903.333             40320.672
##   Exterior2ndMetalSd     Exterior2ndOther   Exterior2ndPlywood
##             3332.311            42598.491             12220.246
##     Exterior2ndStone    Exterior2ndStucco   Exterior2ndVinylSd
##           -30128.014           -15654.765             25997.769
##   Exterior2ndWd Sdng   Exterior2ndWd Shng           GarageCars
##            12333.407             -846.287             10913.402
##           GarageArea           BsmtQualFa           BsmtQualGd
##                4.011           -40123.156            -45696.361
##           BsmtQualTA          TotalBsmtSF           BsmtCondGd
##           -42220.435               15.192              9958.208
##           BsmtCondPo           BsmtCondTA             FullBath
##             4350.833             9677.083              1134.668
##      BsmtFinType1BLQ      BsmtFinType1GLQ      BsmtFinType1LwQ
##            -4660.094             -841.198            -18155.446
##      BsmtFinType1Rec      BsmtFinType1Unf      BsmtFinType2BLQ
##            -9776.562           -18772.189            -26230.900
##      BsmtFinType2GLQ      BsmtFinType2LwQ      BsmtFinType2Rec
##           -12331.307           -22683.091            -22694.504
##      BsmtFinType2Unf
##           -20552.970
```

```
#BsmtQual, GarageCars and TotalBsmtSF are significant, while street became less
#significant.
```

```r
#So in our final model we are using Neighborhood, OverallQual, GrLiveArea,
#GarageCars, BsmtCond, TotalBsmtSF for prediction on our test data.

model_trained <- lm(SalePrice~Neighborhood+BsmtQual+OverallQual+GrLivArea+
                    GarageCars+TotalBsmtSF, data = house_training_data)
#Our model is trained . Now we will predict SalePrice on test dataset
model_trained
```

```
##
## Call:
## lm(formula = SalePrice ~ Neighborhood + BsmtQual + OverallQual +
##     GrLivArea + GarageCars + TotalBsmtSF, data = house_training_data)
##
## Coefficients:
##         (Intercept)  NeighborhoodBlueste   NeighborhoodBrDale
##            11829.85            -15224.19            -20263.60
## NeighborhoodBrkSide  NeighborhoodClearCr  NeighborhoodCollgCr
##             1190.65             33718.60             19040.66
## NeighborhoodCrawfor  NeighborhoodEdwards  NeighborhoodGilbert
##            29941.38             -7562.29             13694.63
##  NeighborhoodIDOTRR  NeighborhoodMeadowV  NeighborhoodMitchel
##           -12619.75             -3292.07               715.54
##   NeighborhoodNAmes  NeighborhoodNPkVill   NeighborhoodNWAmes
##             4474.33             -9460.27              5227.46
## NeighborhoodNoRidge  NeighborhoodNridgHt  NeighborhoodOldTown
##            71869.31             50313.72            -15860.85
##   NeighborhoodSWISU   NeighborhoodSawyer  NeighborhoodSawyerW
##           -12498.42              7636.54             14209.21
## NeighborhoodSomerst  NeighborhoodStoneBr   NeighborhoodTimber
##            23196.85             64829.84             24093.69
## NeighborhoodVeenker           BsmtQualFa           BsmtQualGd
##            49842.35            -50991.51            -46898.76
##          BsmtQualTA          OverallQual            GrLivArea
##           -47116.72             14458.71                45.83
##          GarageCars          TotalBsmtSF
##            12245.75                19.93
```

```r
pred_lm <- predict.lm(model_trained, house_test_data.SalePrice)
house_test_data_with_predictions <- house_test_data.SalePrice %>%
  mutate(predictedSalePrice = pred_lm)
```

Checking accuracy of our model

```r
actual_preds <- data.frame(cbind(actuals=house_test_data.SalePrice$SalePrice,
                               predicteds =  pred_lm))
corelation_accuracy <- cor(actual_preds)
```

Time Series

```r
#timeseries object for Sales Price
salePricets<-ts(actual_preds$predicteds,start=c(2001,1),end=c(2010,12),frequency = 4);
```

```r
#timeseries object for Sales Price and Selling Year
yrSoldts<-ts(house_test_data_with_predictions$YrSold,start=c(2001,1),
             end=c(2010,12),frequency = 4);
salePricets<-ts(house_test_data_with_predictions$predictedSalePrice,
                start=c(2001,1),end=c(2010,12),frequency = 4);

#Checking for frequency data has been collected.
frequency(salePricets);
```

```
## [1] 4
```

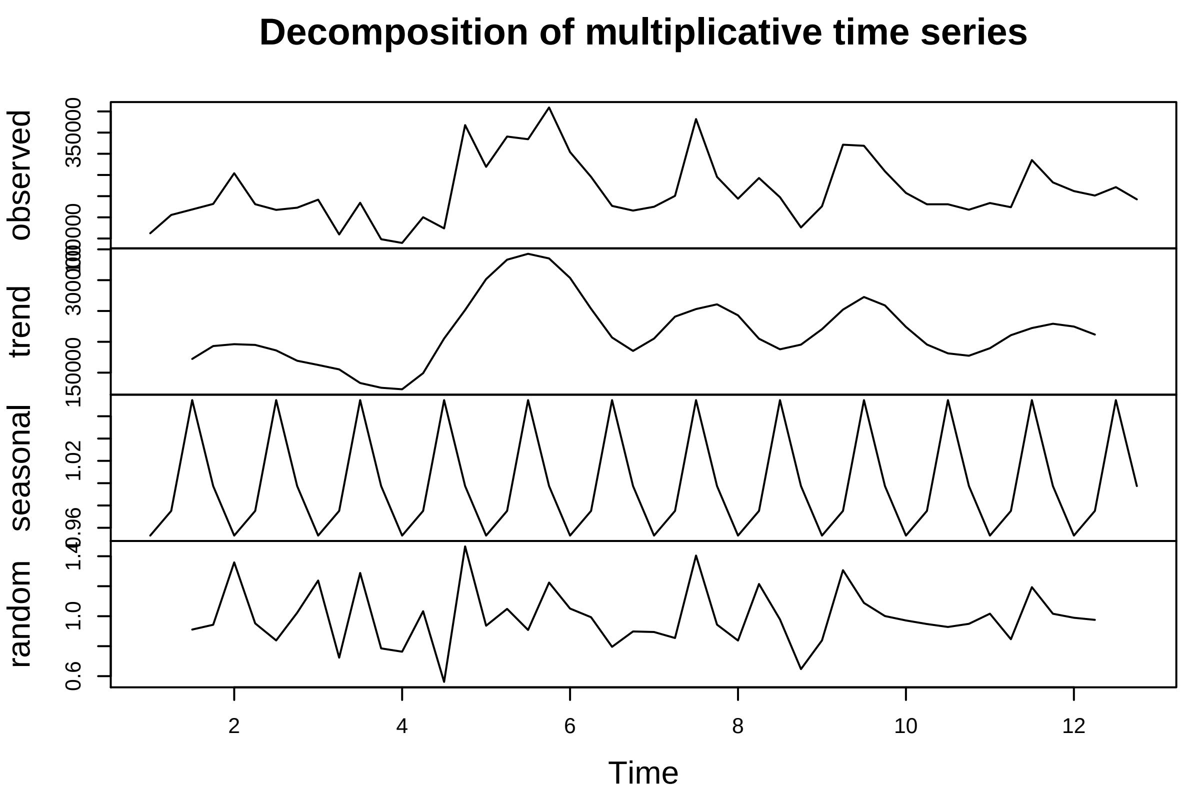
```r
#checking for missing values
sum(is.na(salePricets))
```

```
## [1] 0
```

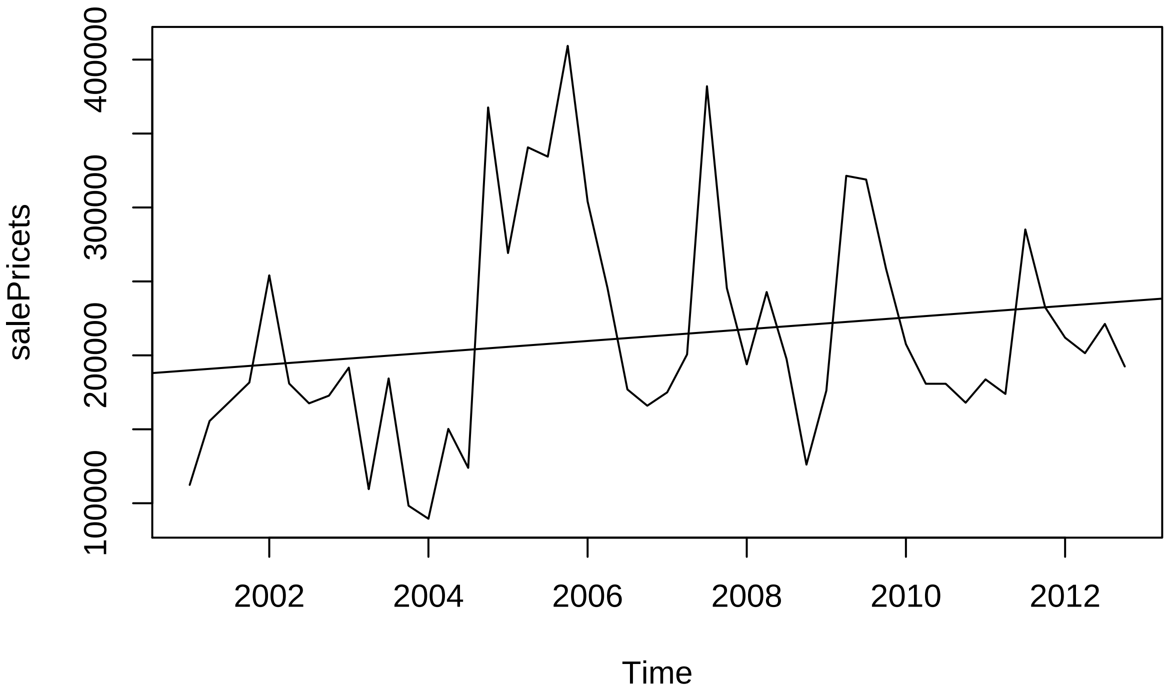
```r
#summary of the data
summary(salePricets)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   89503  171699  192087  213181  247655  409304
```

```r
#decomposing the data into trend, seasonal, regular and random components
tsdata<-ts(salePricets,frequency = 4)
ddata<-decompose(tsdata,"multiplicative")
plot(ddata)
```



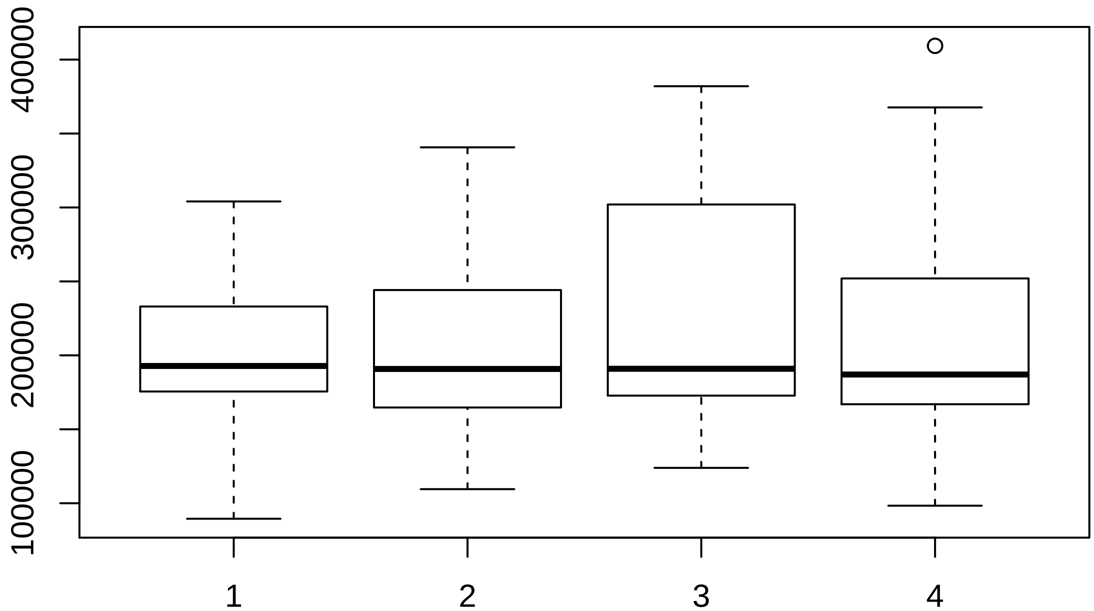

```r
#checking the original trend in data while performing linear regression.
plot(salePricets)
abline(reg=lm(salePricets~time(salePricets)))
```




```r
cycle(salePricets)
```

```
##      Qtr1 Qtr2 Qtr3 Qtr4
## 2001    1    2    3    4
## 2002    1    2    3    4
## 2003    1    2    3    4
## 2004    1    2    3    4
## 2005    1    2    3    4
## 2006    1    2    3    4
## 2007    1    2    3    4
## 2008    1    2    3    4
## 2009    1    2    3    4
## 2010    1    2    3    4
## 2011    1    2    3    4
## 2012    1    2    3    4
```

```r
#boxplot for quaterly data to analyse in which quater sales price is going up
boxplot(salePricets ~cycle(salePricets, xlab="Date"))
```




```r
#checking for the best model
priceModel<-auto.arima(salePricets)
priceModel
```

```
## Series: salePricets
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1       mean
##       0.5183  210572.87
## s.e.  0.1237   18628.69
##
## sigma^2 estimated as 4.196e+09:  log likelihood=-599.02
## AIC=1204.04   AICc=1204.59   BIC=1209.66
```

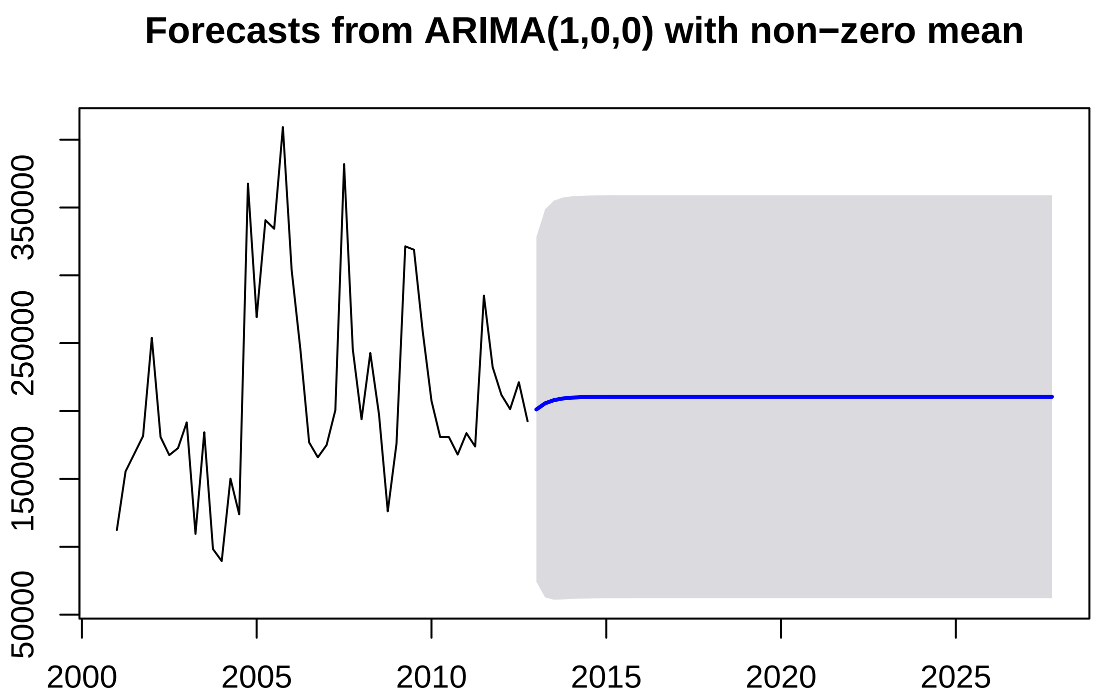
```r
#running with trace to compare the information criterion
auto.arima(salePricets,ic="aic",trace= TRUE)
```

```
##
##  ARIMA(2,0,2)(1,0,1)[4] with non-zero mean : Inf
##  ARIMA(0,0,0)           with non-zero mean : 1216.753
##  ARIMA(1,0,0)(1,0,0)[4] with non-zero mean : 1205.981
##  ARIMA(0,0,1)(0,0,1)[4] with non-zero mean : 1210.273
##  ARIMA(0,0,0)           with zero mean     : 1321.611
##  ARIMA(1,0,0)           with non-zero mean : 1204.044
##  ARIMA(1,0,0)(0,0,1)[4] with non-zero mean : 1205.977
##  ARIMA(1,0,0)(1,0,1)[4] with non-zero mean : Inf
##  ARIMA(2,0,0)           with non-zero mean : 1205.96
##  ARIMA(1,0,1)           with non-zero mean : 1205.992
##  ARIMA(2,0,1)           with non-zero mean : 1207.638
##  ARIMA(1,0,0)           with zero mean     : 1215.511
##
##  Best model: ARIMA(1,0,0)           with non-zero mean
```

```
## Series: salePricets
## ARIMA(1,0,0) with non-zero mean
```

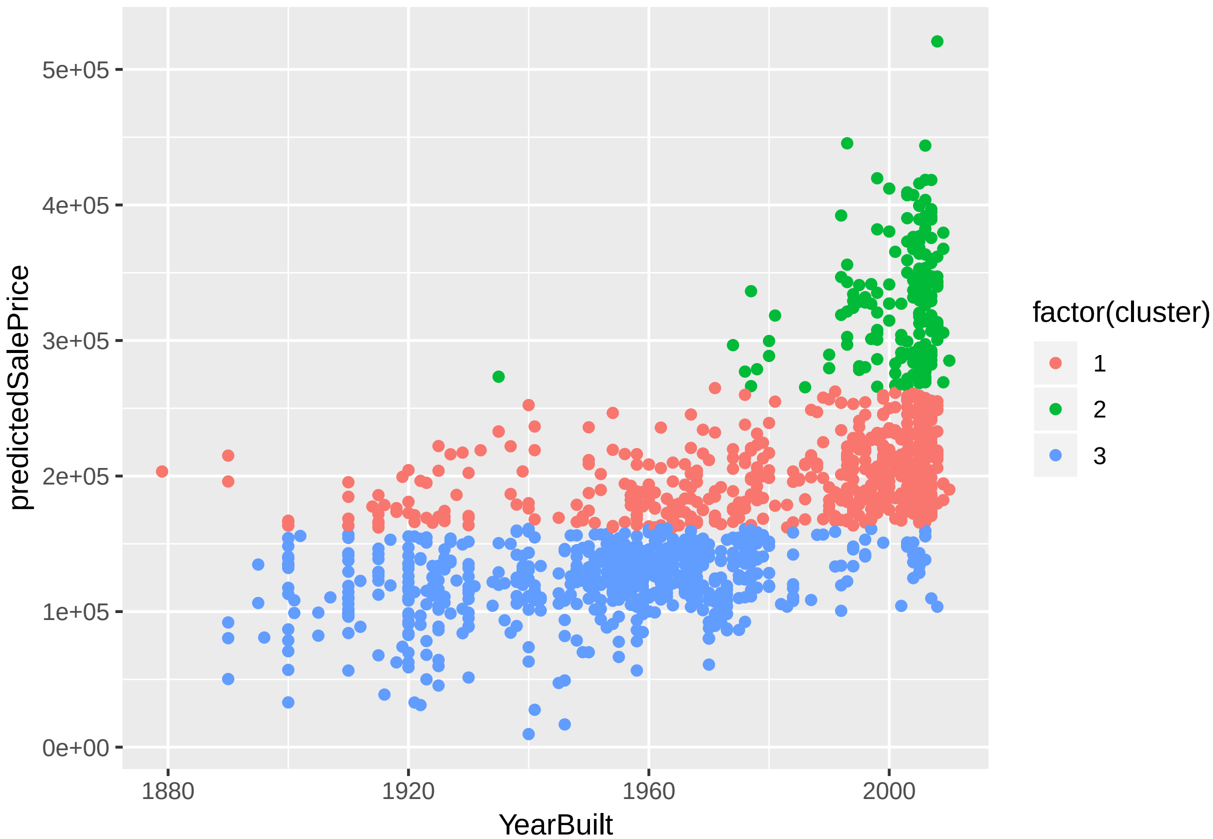
```
##
## Coefficients:
##          ar1       mean
##       0.5183  210572.87
## s.e.  0.1237   18628.69
##
## sigma^2 estimated as 4.196e+09:  log likelihood=-599.02
## AIC=1204.04   AICc=1204.59   BIC=1209.66
```

```r
#Using the model to forecast for next 5 years with 95% accuracy
priceForecast<-forecast(priceModel,level=c(95),h=5*12)
plot(priceForecast)
```

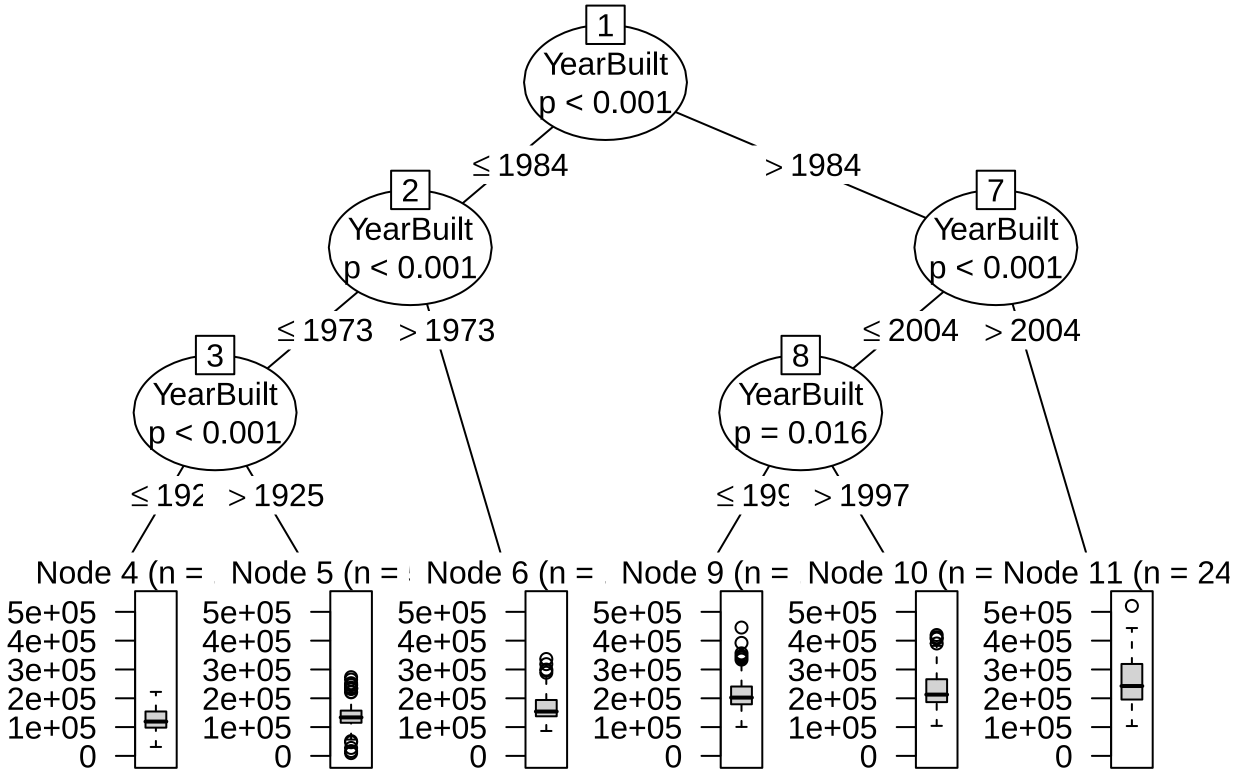


Clustering

```r
# Get Predicated Sale Price with Year Built
sale_price_with_built_year <- house_test_data_with_predictions %>%
  select(YearBuilt, predictedSalePrice) %>% na.omit()

cluster <- kmeans(sale_price_with_built_year, 3)$cluster
cbind(sale_price_with_built_year, cluster) %>%
  ggplot((aes(x = YearBuilt, y = predictedSalePrice, color = factor(cluster)))) +
  geom_point()
```




```
tree <- ctree(predictedSalePrice ~ ., data = sale_price_with_built_year,
controls = ctree_control(minbucket = 100))
plot(tree)
```



Conclusion

We saw that the variables which we used to build our linear model were effecting the sale price such as Neighborhood, BsmtQual, OverallQual, GrLivArea, GarageCars, TotalBsmtSF. Then we used Time Series to forecast sale prices for the next 10 years. In the end we saw by applying k-means clustering that house prices with respect to the year they were built in can be clustered into high, low and mid sale prices. We can see that the most expensive houses can be found after 1980(year built) (approx).