How do you predict your property price?

Real estate speicalist Minsoo Seok

09/05/2021

Abstract

It will be very important for real estate angency to understand the pattern of sale price and predict the sale price based on pattern becasue it impacts on their marketing strategies. This report follows data science methodology in order to present evidence for the sale price prediction. This report manily try to find elements which affects properties and the elements will be used for linear regression. This report found the key 10 elements such as location, overall quaility of property. It helps predict properies sale price and It will be main information for real estate agency to organize their marketing strategies.

Set Working directory

```
setwd("~/Desktop/EDA/FINAL PROJECT")
```

loading Packages used for project

```
library(dplyr)
library(tidyverse)
library(ggplot2)
library(GGally)
library(ggrepel)
library(egg)
library(reshape2)
library(corrplot)
library(lubridate)
library(modelr)
library(car)
```

Dataset used for project

```
train <- read.csv("train.csv")
test  <- read.csv("test.csv")

dim(train)
dim(test)
head(train)
head(test)
colnames(train)
colnames(train)
```

1.PROBLEM IDENTIFICATION

You are a real estate analyst. it is always important to predict saleprice of properties by investigating elements which affect saleprice. The dataset provided include information of saleprice and elements impact to the saleprice.

This report uses professional structure which follows the data science methodology including preprocessing, EDA, linearmodel, etc,..

You are setting the 7 important questions regarding the dataset and analyise the questions to predict the saleprice of properties. You will mainly use data analysis techniques and linear regression model.

There are two datasets privded "train" and "test". The train dataset consist of 81 variables with 1460 observation. and test dataset consist of 81 variables with 1459 observation. You will mainly use train dataset and test dataset will be used only for "Evaluation" the model. The root mean squared Error(RMSE) is used for computation of the model error.

1) What is the distribution of sale price over the years? if there are outliers, what is the outcome without outliers? will you use median price or mean price?

-This will be main question for every real estate agency. It is quite important to analyze what is the distribution of sale price so that the price pattern will be found and it helps to understand any information regarding price.

2) Is there any seasonal, year and month impact on amount of properties on salep rice and median price?

-This question helps to identify if there is pattern depends on period. This information is significant for real estate agency to make better marketing strategy.

3) Which classification of the sale are most and is there any impact on saleprice?

-This question is to understand if there is impact of classification. If any pattern exist for this column, it will be helpful to know predict the sale price. For instance, agriculture area is normally cheaper than commercial area.

4) Which Neighborhoods records the high sale price and low sale price compared to each others? where is the top 5 median price Neighborhoods and how the plot looks like in time series?

-This question is one of the most important questions. Location is always key element of property. There must be pattern and it will help to predict future sale price.

5) Does overallquall has a great impact on sale price and is there any specific pattern?

-It is obvious if the quality of property is higher, the sale price will be higher too. As well as, it can be found that how much the quality affects to sale price. (such as 1 quality higher, the price increase —)

6) Which SaleType is most common and does SaleType has an impact on median salep rice?

-It can be found what is the trend way to purchase properties through this question. It is important to know the trend so that it can be understood that common way to secure funds.

7) Does Year built has an impact on sale price?

-It is well known that newly built house is more expensive because there are more chances to be modern style. It will be found that if there might be a pattern or not through this question.

8) Which elements have an great impact on sale price?

-Above questions are not enough to check all the elements which affects sale price. So, this question will be very important to analyze linear model regression regarding sale price

explanation of variables

MSSubClass: Identifies the type of dwelling involved in the sale.

MSZoning: Identifies the general zoning classification of the sale.

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

LotShape: General shape of property

Neighborhood: Physical locations within Ames city limits

OverallQual: Rates the overall material and finish of the house

OverallCond: Rates the overall condition of the house

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

GrLivArea: Above grade (ground) living area square feet

TotalBsmtSF: Total square feet of basement area

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

FireplaceQu: Fireplace quality

GarageType: Garage location

GarageYrBlt: Year garage was built

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

GarageCond: Garage condition

PoolArea: Pool area in square feet

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

2.DATA PREPROCESSING

It is very important to preprocess your data before you actually use it to get insight. clean data helps avoid getting any error or misunderstanding.

Dealing with missing values

```
sort(colSums(is.na(train)))
```

Check how many missing values(NA) are in each column Columns which have more than 30% NA values are PoolQC, MiscFeature, Alley, Fence and FireplaceQu.

Remove columns which have more than 30%

```
train<-train[, colMeans(is.na(train)) <= 0.3]
```

The code detects columns which have more than 30% NA values and remove the columns removing columns is because it does not have enough data on these column to compare to other values.

```
sort(colSums(is.na(train)))
```

Columns which have NAs: 3(Numerical) , 11(Categorical)

Numerical:LotFrontage,GarageYrBlt,MasVnrArea

Categorical:GarageCond,GarageQual,GarageFinish,GarageType,BsmtFinType2,BsmtExposure,BsmtFinType1,BsmtCond,Bs

It should be divided first because it should be dealt with in a differnt way.

Check the ditribution of numerical dataset which have NA values

```
ggplot(train,aes(x=LotFrontage))+
geom_histogram()
```

```
ggplot(train,aes(x=GarageYrBlt))+
geom_histogram()
```

```
ggplot(train,aes(x=MasVnrArea))+
geom_histogram()
```

Those Three numerical values have skewness obviously, so median value will be used for NA values.

Imputing NA values in numerical column to Median values

```
train <- train %>%
  mutate_if(is.numeric, ~replace_na(.,median(., na.rm=TRUE)))
```

Code detects numerical columns and change NA values to median.

Imputing Na values in categorical column to Mode values

```
Mode <- function(x) {
  ux <- na.omit(unique(x) )
 tab <- tabulate(match(x, ux)); ux[tab == max(tab) ]
}

sapply(train,Mode)

train$GarageCond[is.na(train$GarageCond)] <- "TA"
train$GarageQual[is.na(train$GarageQual)] <- "TA"
train$GarageFinish[is.na(train$GarageFinish)] <- "Unf"
train$GarageType[is.na(train$GarageType)] <- "Attchd"
train$BsmtFinType2[is.na(train$BsmtFinType2)] <- "Unf"
train$BsmtExposure[is.na(train$BsmtExposure)] <- "No"
train$BsmtFinType1[is.na(train$BsmtFinType1)] <- "Unf"
train$BsmtCond[is.na(train$BsmtCond)] <- "TA"
train$BsmtQual[is.na(train$BsmtQual)] <- "TA"
train$MasVnrType[is.na(train$MasVnrType)] <- "None"
train$Electrical[is.na(train$Electrical)] <- "SBrkr"
```

NA values are changed to "Mode" in every categorical column. Program R does not provide mode fuction so Mode is defined first in the code and detect what is the code for each column with sapply, and it is applied into each categorical columns which have Na values.

Check if there is still Na values in the dataset

```
sum(is.na(train))
```

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

There is no NA values anymore in any column. So the dataset is ready to EDA part!

Corelation heatmap(numerical column)

We can see overview simply how much numerical columns related to saleprice throughout corelation heatmap. There are a lot of numerical columns so top 10 variables is picked to plot corelation heatmap.

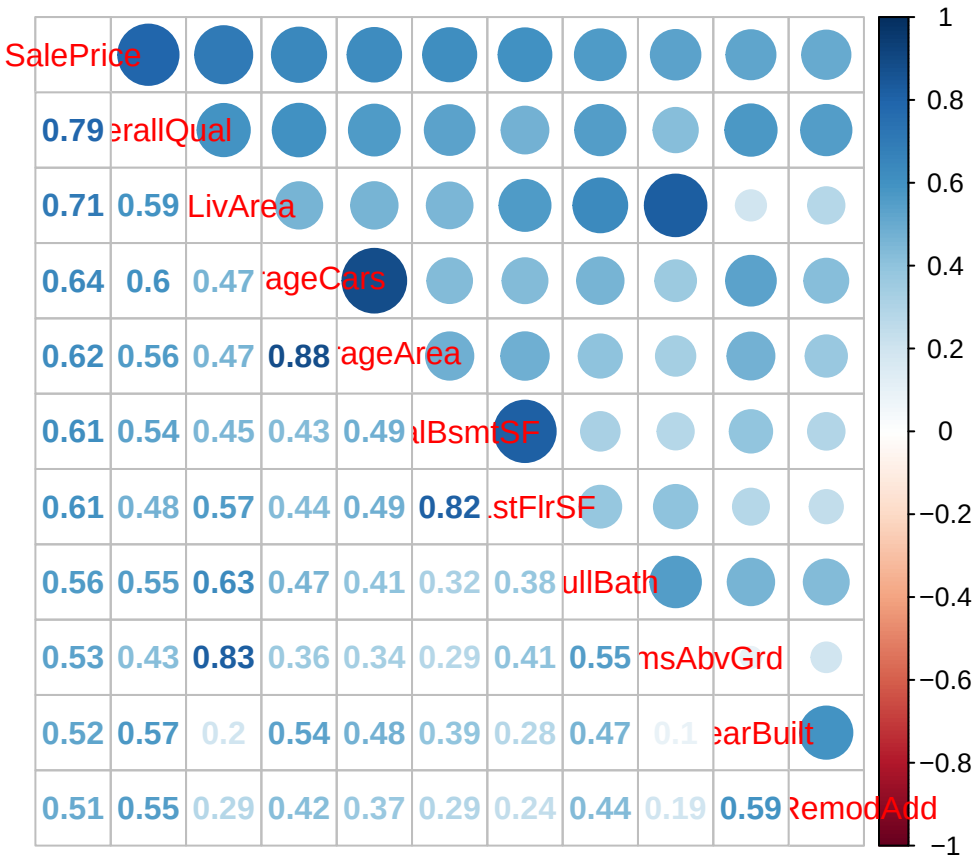
```
Numeric_column<- train %>%select(where(is.numeric))
cor_Numeric<-round(cor(Numeric_column),5)

Saleprice_sort <-as.matrix(sort(cor_Numeric[,'SalePrice'], decreasing = TRUE))

top10 <-names(which(apply(Saleprice_sort, 1, function(x) abs(x)>0.507)))

numeric_data <- cor_Numeric[top10, top10]

corrplot.mixed(numeric_data)
```



The heatmap cleary shows numerical columns which have greater impact on saleprice. Top 10 numerical values which record high corelation are OverallQual,GrLivArea,GarageCars,GarageArea,TotalBsmtSF,X1stFlrSF,FullBath,TotR
Therefore, these numerical columns are key for further analysis.

3.EXPLORATORY DATA ANALYSIS AND VISUALISATION

Price range distribution

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

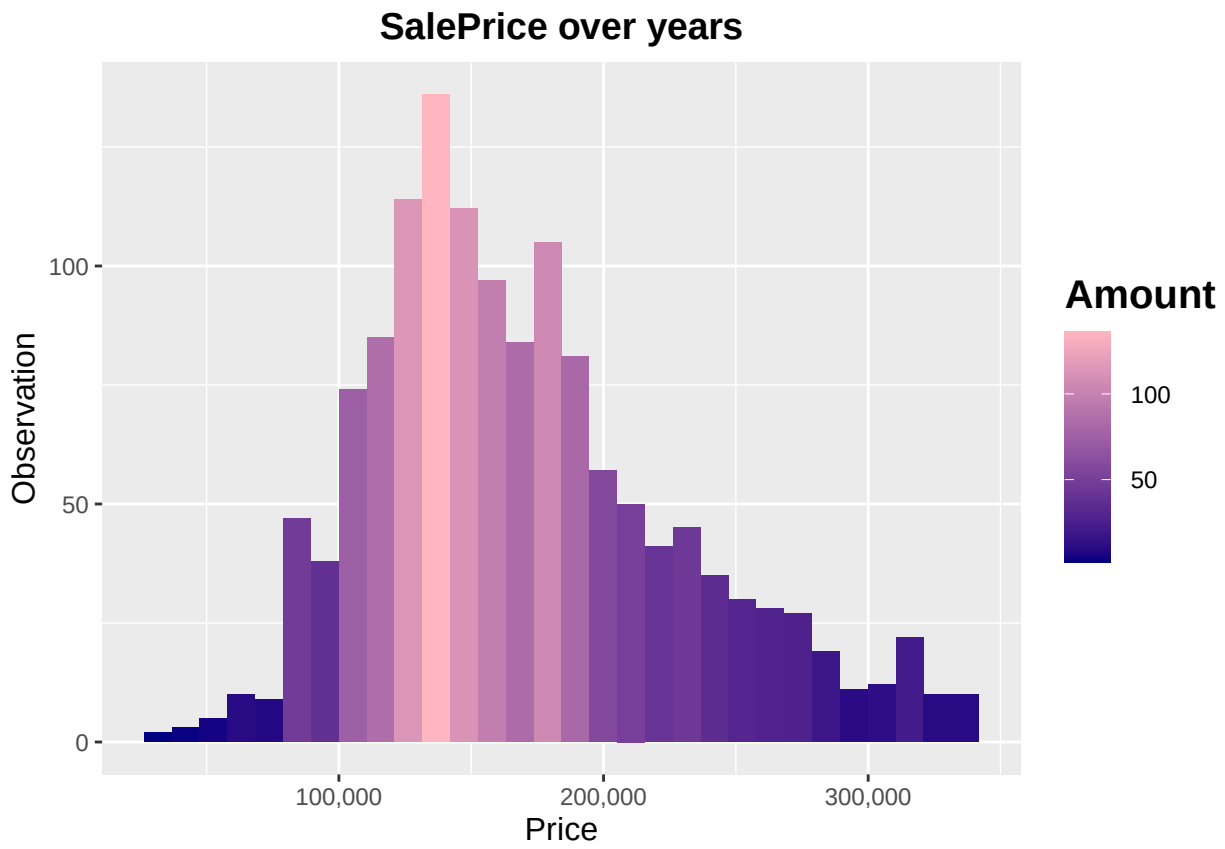
median price distribution of overall price is skewed to right. The mean price is higher than median price.It can be assumed that some extreme expensive houses (such as 755,000)are outliers and they affect the mean value a lot. Statistical parameters are highly sensitive to outliers. Therefore, it would be great to see the outcome once again without outliers to extract insight of the dataset.

Price range distribution without outliers

```
##  [1] 345000 385000 438780 383970 372402 412500 501837 475000 386250 403000
## [11] 415298 360000 375000 342643 354000 377426 437154 394432 426000 555000
## [21] 440000 380000 374000 430000 402861 446261 369900 451950 359100 345000
## [31] 370878 350000 402000 423000 372500 392000 755000 361919 341000 538000
## [41] 395000 485000 582933 385000 350000 611657 395192 348000 556581 424870
## [51] 625000 392500 745000 367294 465000 378500 381000 410000 466500 377500
## [61] 394617
```

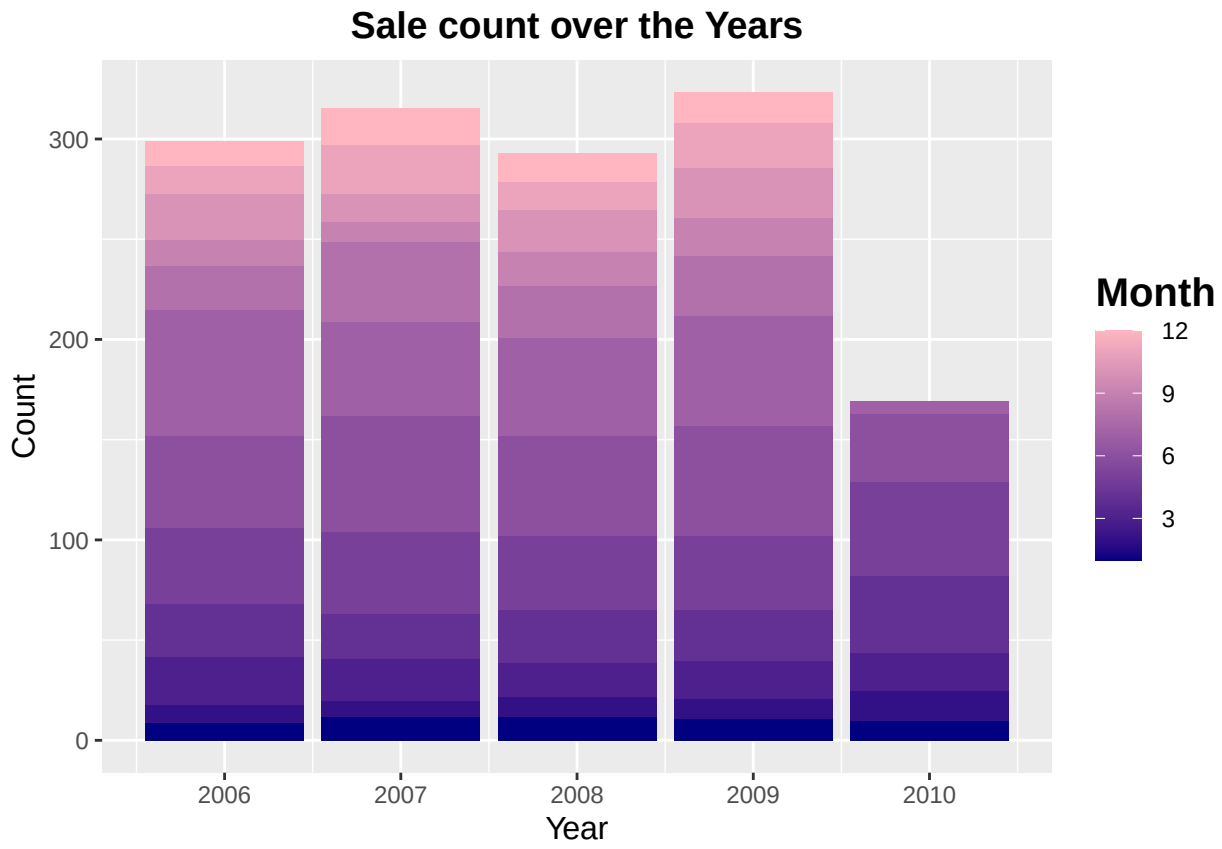
```
##  [1]   12   54   59  113  152  162  179  186  225  232  279  310  314  321  322
## [16]  337  350  379  390  441  474  478  482  497  516  528  586  592  609  643
## [31]  645  655  662  665  679  689  692  703  719  770  775  799  804  826  878
## [46]  899  988  991 1047 1143 1170 1182 1183 1229 1244 1268 1269 1354 1374 1389
## [61] 1438
```

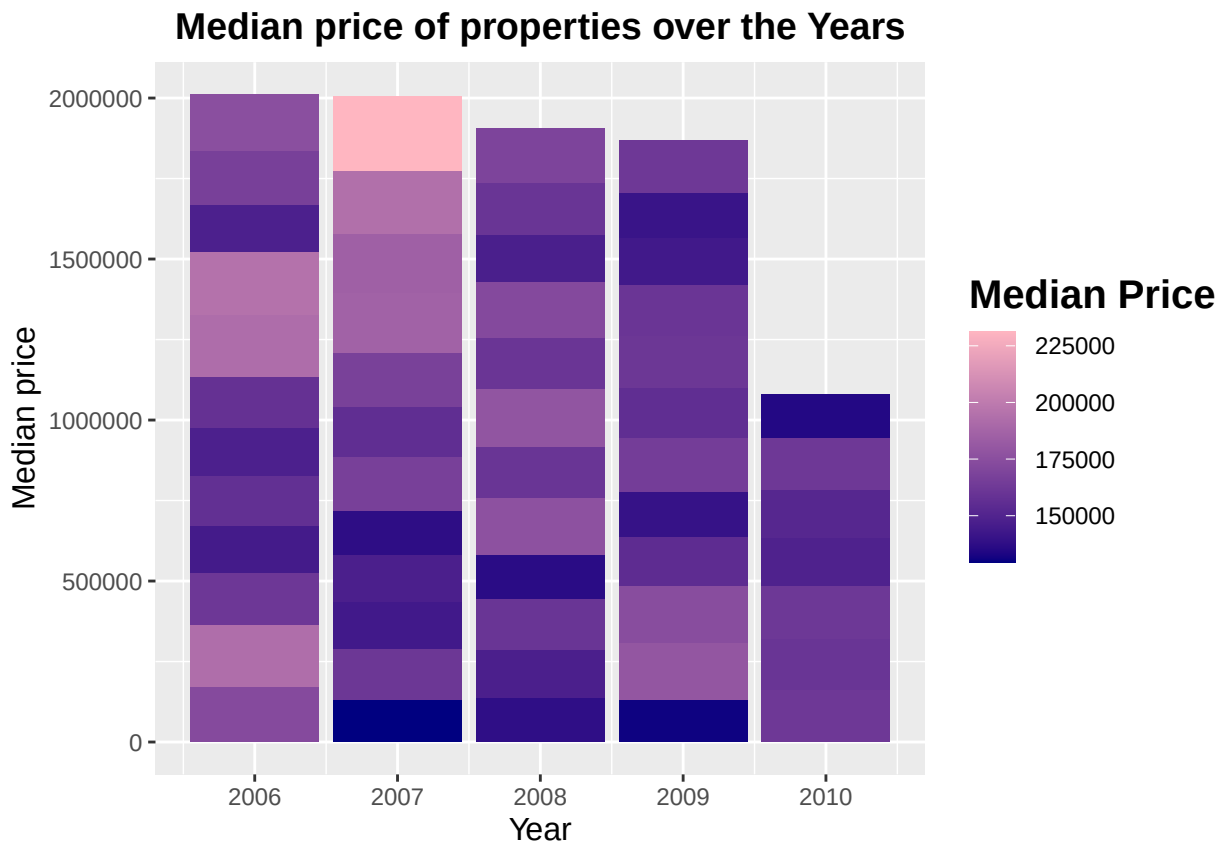
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129000  159500  170237  203500  340000
```



Some of extrem expensive houses are outliers. The max price is 340000 without outliers(it was 755000). We can check the median and mean decrease around $4500, $10000 each. It still shows skewness, therefore, it will be better to use median price rather than mean price.

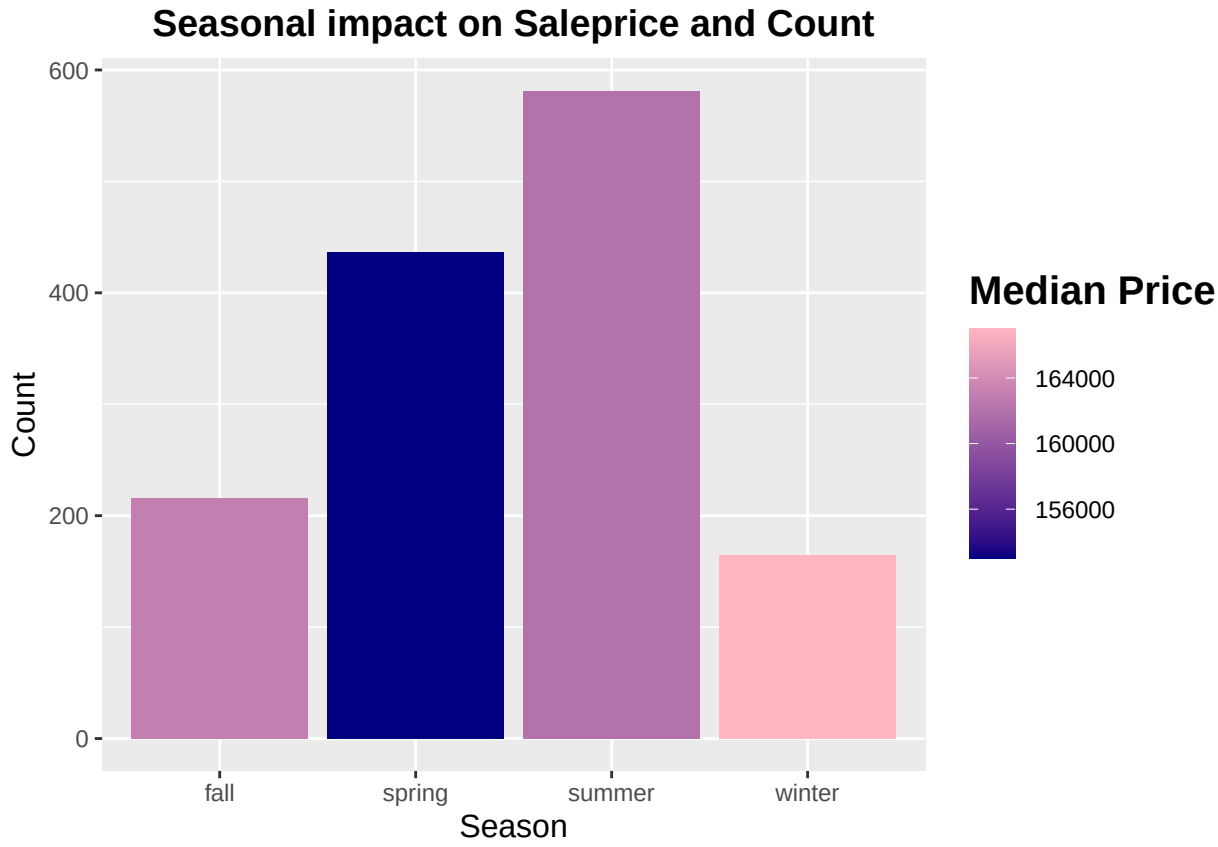
Year / Month house sold distribution





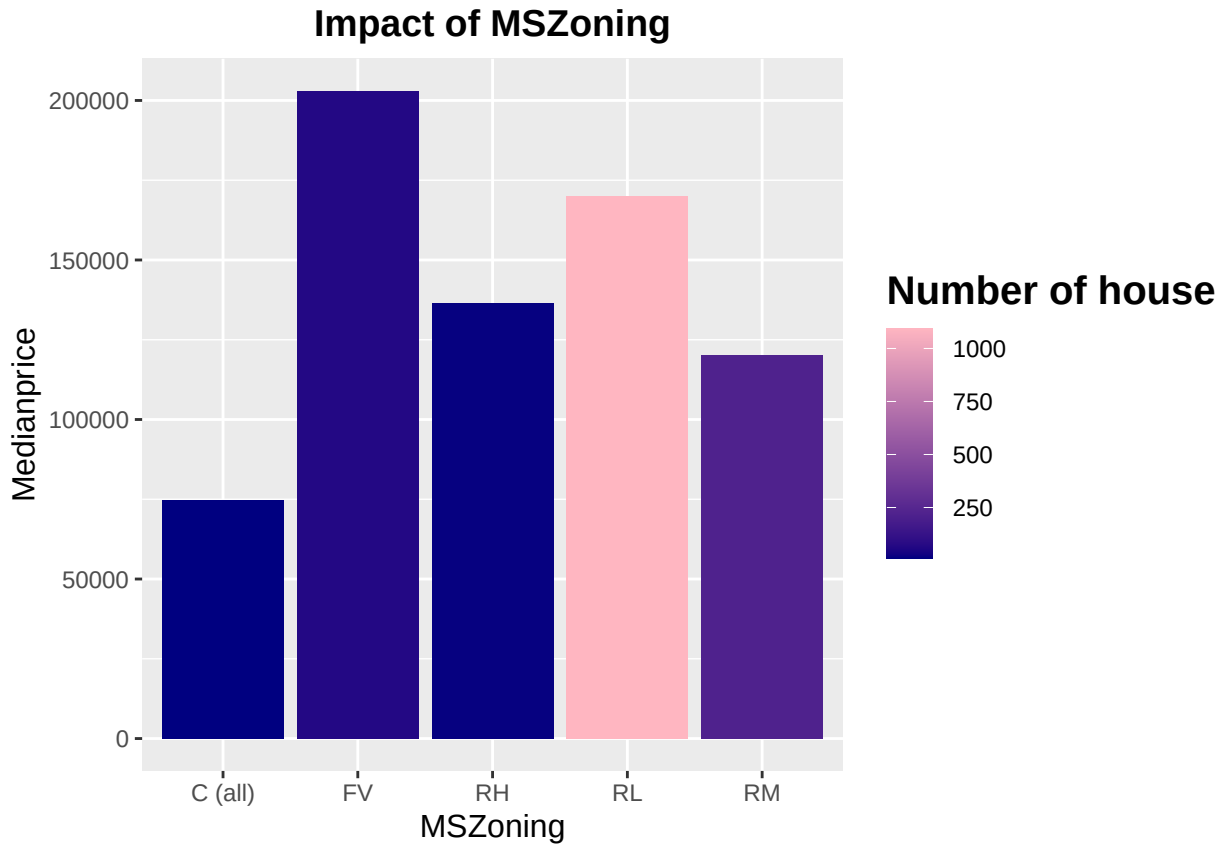
The highest housesold recorded in 2009. More than 350 houses sold this year. 2010 record the lowest housesold among these year. There is no specific pattern of housesold based on year. May,June,July records high housesold compared to other months. it is nearly symetric distribution. From feburary, it increase gradually until June and it drops to december. interestingly, The median price over the given years(2006 to 2010) is getting smaller.

Seasonal impacts on median price and number of properties sold



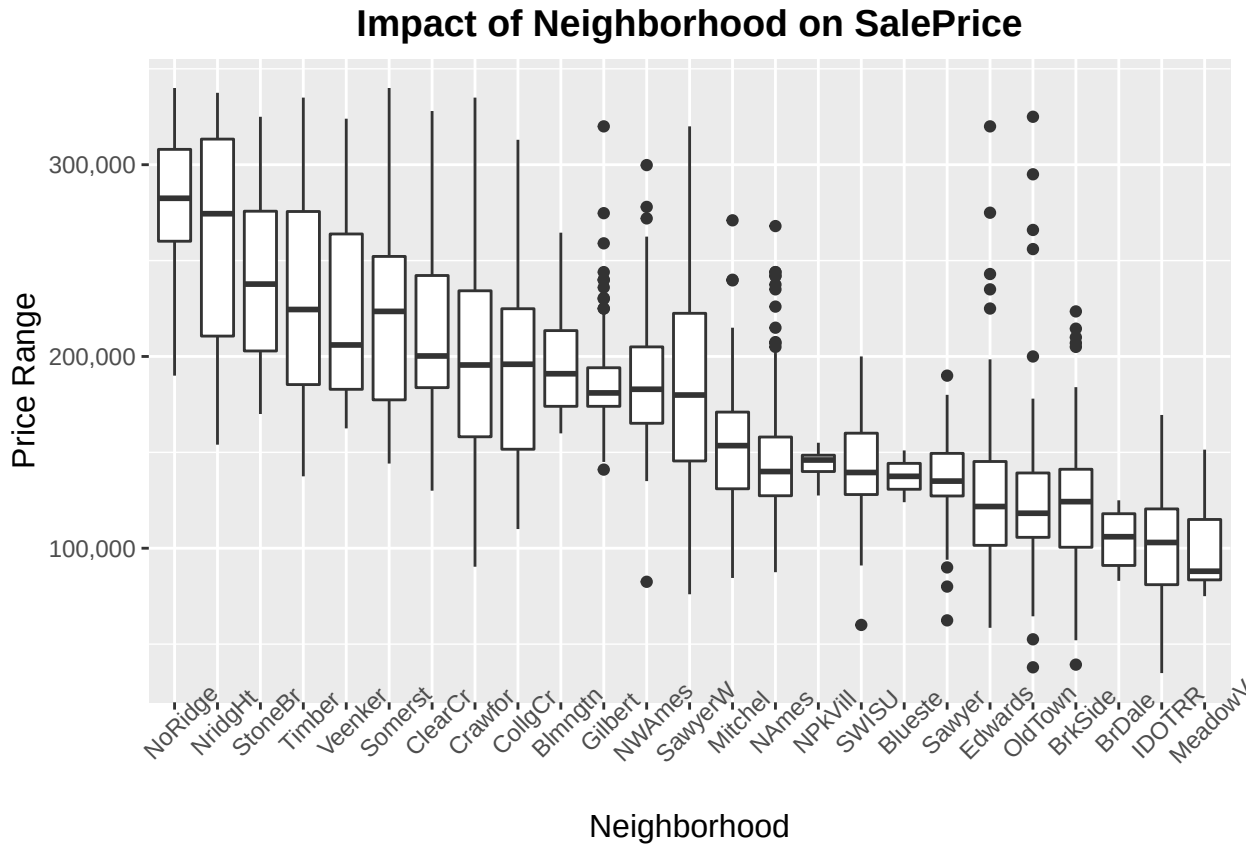
There are seasonal impact on amount of sale. The amount of sale records higher spring and summer(3~8) compared to other period over the year. Especially the amount of sale is the highest in summer which record right under 600. The amount of sale is the least in winter which record less than 200 over the given year. It can be assume that the weather based on season affects the amount of sale. However, It can not be found specific patter in saleprice depends on season.

MSZoning impacts on median price and number of properties sold



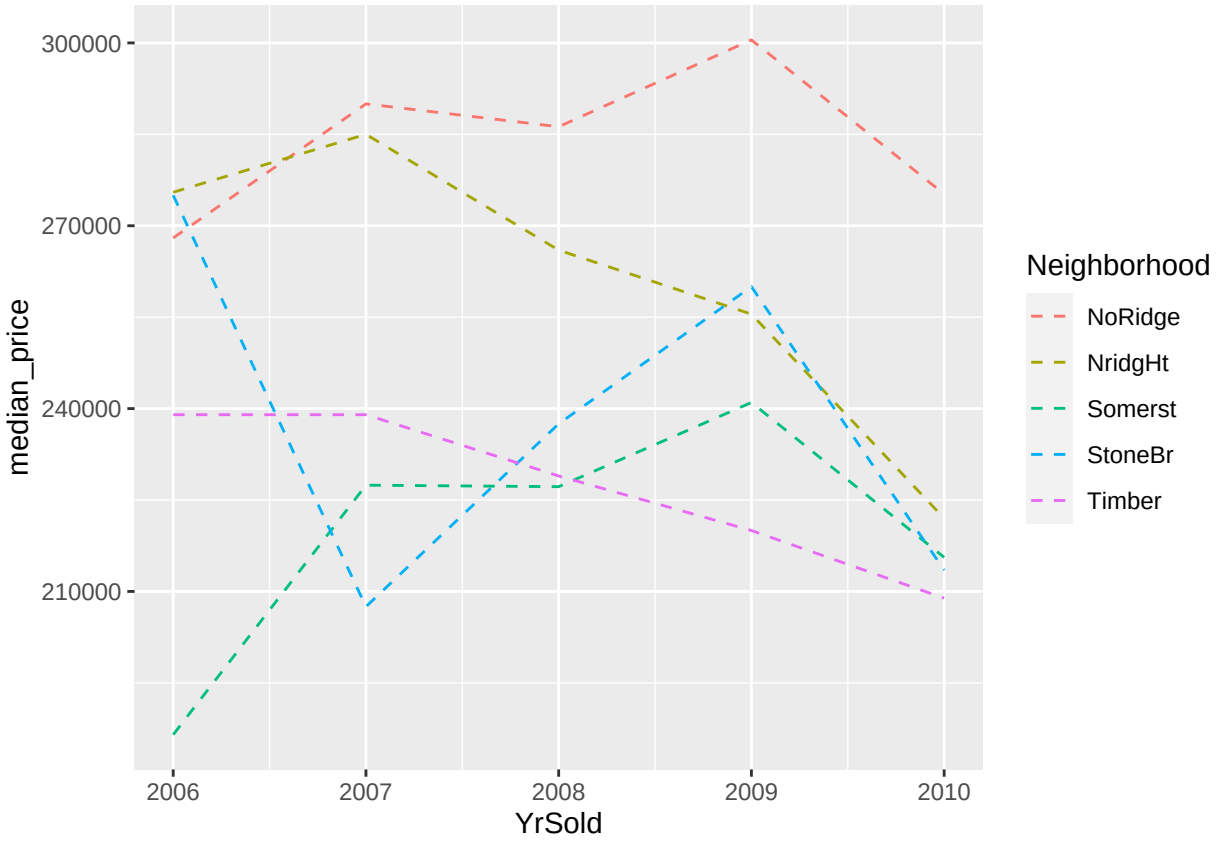
Most of cases are RL(Residential low Density) which records more than 1000 out of 1400 observation. This RL zone records 170,000 median price. Floating Village Residential records the highest median price and commercial area record the least median price among the categories

Neighborhood and sale price distrubution



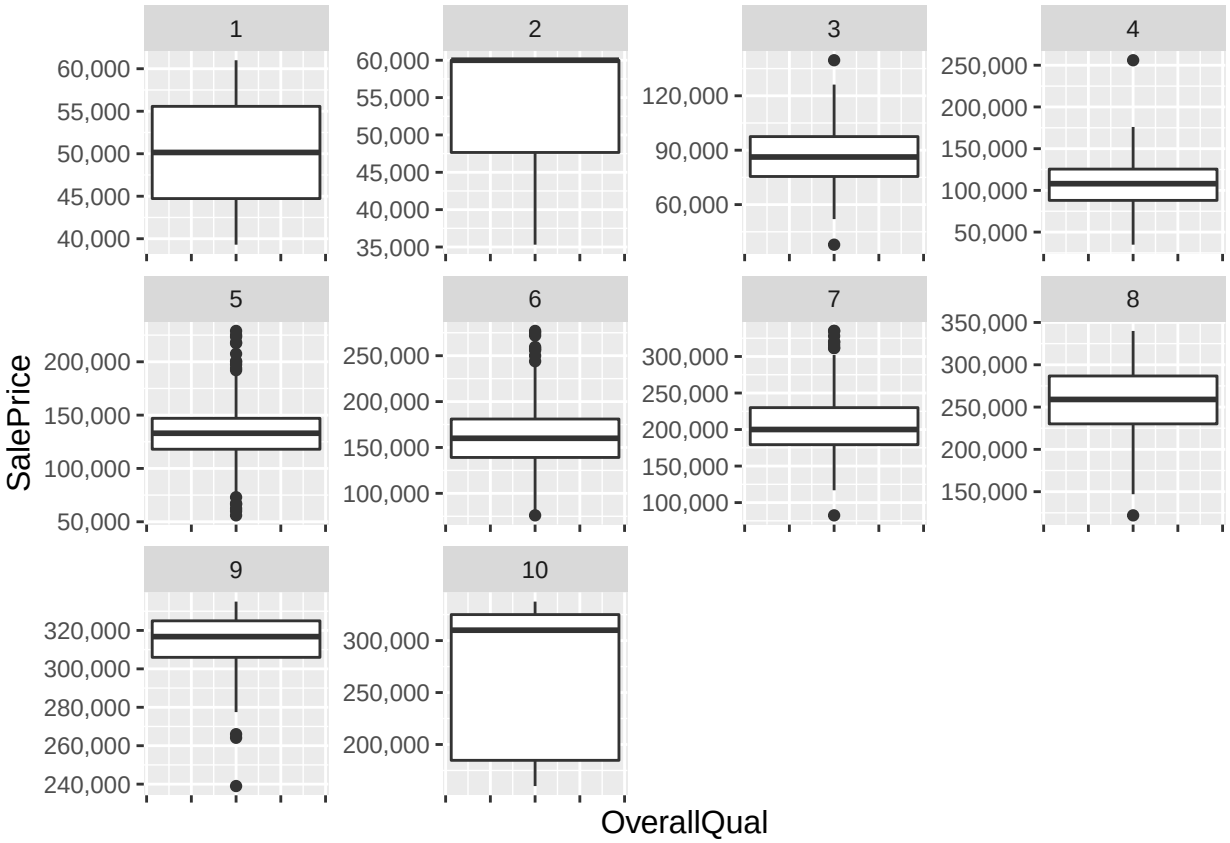
According to the box graph above, NridgHt records the highest price among neighborhood and NoRidge is the second highest. Only These two neighborhood record more than $250,000 as a median price. The Meaddow records the least median price around $100,000. Therefore, it can tell that the highest price of neighborhood records more than 2.5 times to the least price of neighborhood.

TOP5 Median price neighborhood and median price based on time series plot



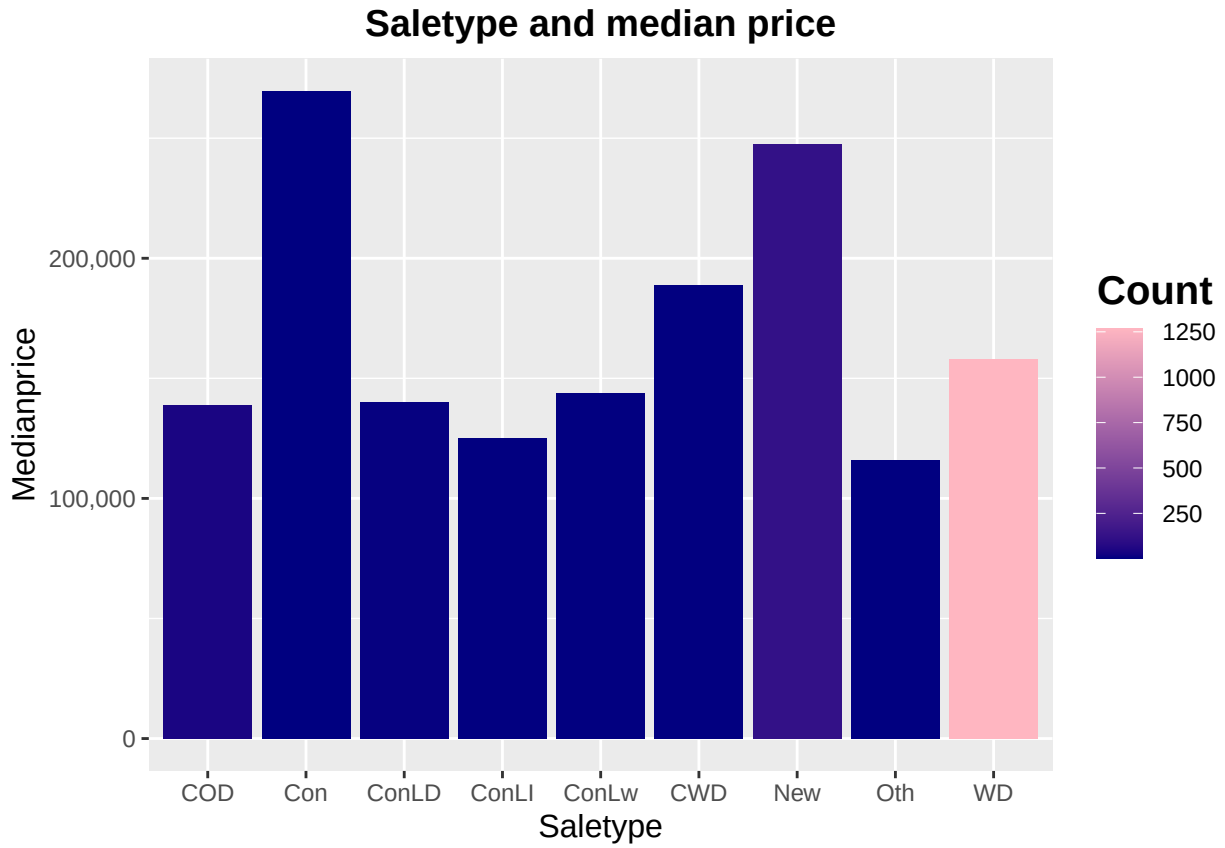
The top 5 neighborhood which record high median price are NoRidge, NridgHt, StoneBr, Timber and Somerst. It is investigated that how their median price is changed based on year. There is no specific pattern for 5 neighborhood of their median price. but ,mainly, they record median price between $210,000 to $290,000

OverallQual and sale price distrubution



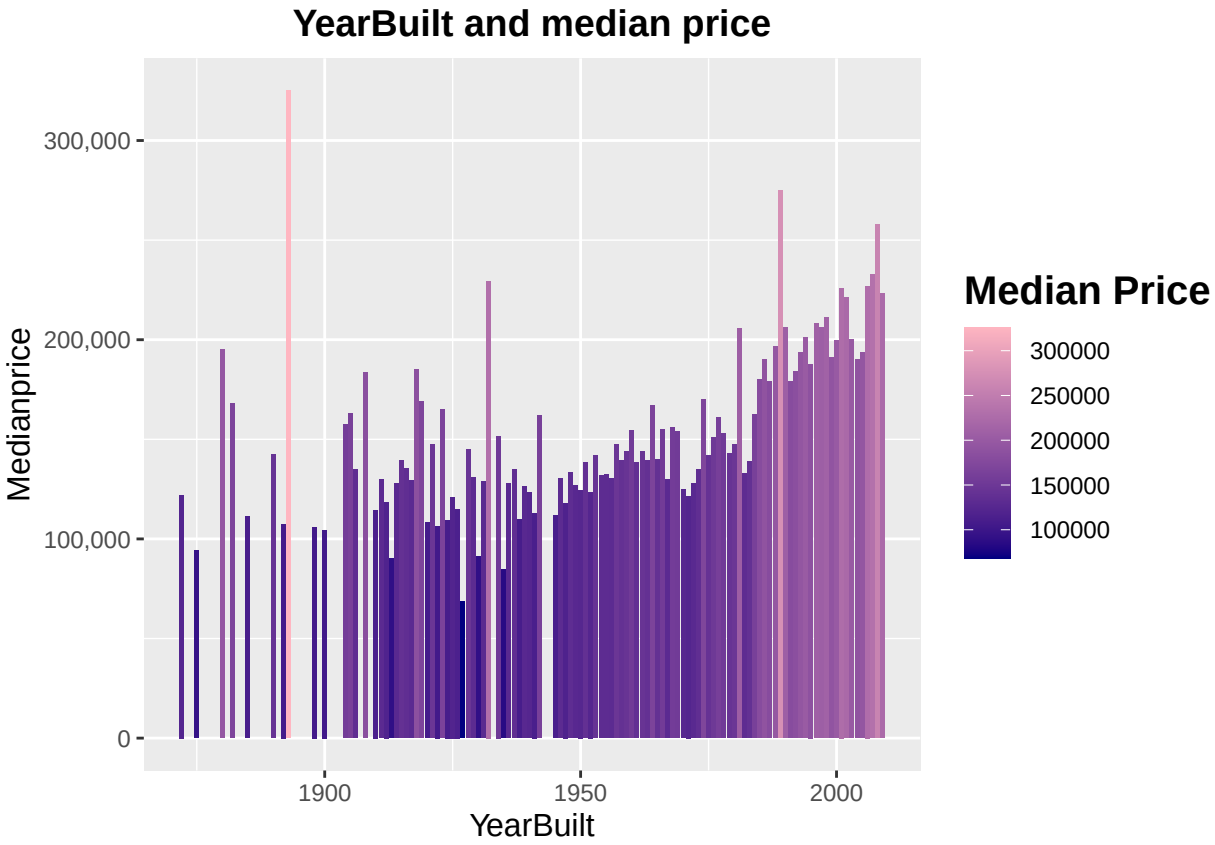
There is vary clear relationship between overallqual and saleprice.The pattern is 1 overallqual is higher , $150,000 is higher on average. There are several outliers on each Quality. Quality 2 and 10 have extreme skewness to left.

SaleType and sale price distrubution



Most of saletype is warranty deed-conventional. it records 1267 out of 1460. It is a trend way to buy properties. "Home just constructed and sold" was second common way to sell property which record 122 out of 1460. The median price of Contract 15% Down payment regular terms was the highest which record 269600$. It can be assumed that buyer pay cash for it.

YearBuilt and sale price distrubution



In terms of amount of sale, it is shown that properties built recently have bigger sale number of properties compared to older properties. interestingly, property records the higest saleprice is builtbefore 1900. It can be assumed that this property has a historic value. overall, the median price is higher when it is built recently,however it does not happen always.

```
plot(train_clean$MSSubClass, train_clean$SalePrice)
```

```
plot(train_clean$LotFrontage, train_clean$SalePrice)
```

```
plot(train_clean$LotArea, train_clean$SalePrice)
```

```
plot(train_clean$YearRemodAdd, train_clean$SalePrice)
```

```
plot(train_clean$LotArea,train_clean$SalePrice )
```

```
plot(train_clean$OverallCond,train_clean$SalePrice)
```

```
plot(train_clean$Bedroom,train_clean$SalePrice)
```

More candidates of variables are tested in order to find corelation. It can be found that the pattern exists "YearRmodAdd" "overallcond" and "LotArea" variables. This variable will be used for linear model.

4.FURTHER PREPROCESSING

```
selected_train<-train_clean %>% select(c(SalePrice,YearBuilt,YearRemodAdd,
LotArea,OverallCond,OverallQual,GrLivArea,GarageCars,
GarageArea,TotalBsmtSF,X1stFlrSF,
FullBath,MSZoning ,SaleType,Neighborhood))
```

Finally 14 columns are chosen for linear modelbased on correlation heat map and EDA activities. These columns shows certain pattern even though it is small pattern toward sale price.14 columns consist of numeric and character.

5.MODELLING

```
linear_model <- lm(log(SalePrice)~YearBuilt+YearRemodAdd+LotArea+OverallCond+OverallQual + GrLivArea +
                     GarageCars + GarageArea + TotalBsmtSF +
                   X1stFlrSF +FullBath +MSZoning + SaleType +
                  Neighborhood, selected_train)

rmse(linear_model, test)
```

```
## [1] 0.1479232
```

The first linear model records 0.1479232. More formulation will be tested to minimize RMSE.

```
summary(linear_model)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ YearBuilt + YearRemodAdd + LotArea +
##     OverallCond + OverallQual + GrLivArea + GarageCars + GarageArea +
##     TotalBsmtSF + X1stFlrSF + FullBath + MSZoning + SaleType +
##     Neighborhood, data = selected_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76339 -0.05991  0.00868  0.07526  0.44706
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.821e+00  6.903e-01   5.535 3.73e-08 ***
## YearBuilt            2.389e-03  3.093e-04   7.722 2.21e-14 ***
## YearRemodAdd         8.701e-04  2.742e-04   3.174 0.001540 **
## LotArea              1.499e-06  5.394e-07   2.779 0.005527 **
## OverallCond          5.252e-02  4.275e-03  12.285  < 2e-16 ***
## OverallQual          6.718e-02  5.043e-03  13.320  < 2e-16 ***
## GrLivArea            2.180e-04  1.305e-05  16.712  < 2e-16 ***
## GarageCars           6.953e-02  1.224e-02   5.679 1.66e-08 ***
## GarageArea           3.037e-05  4.234e-05   0.717 0.473335
```

```
## TotalBsmtSF          8.874e-05  1.632e-05   5.437 6.41e-08 ***
## X1stFlrSF            1.615e-06  2.015e-05   0.080 0.936104
## FullBath             7.252e-03  1.066e-02   0.680 0.496499
## MSZoningFV           3.863e-01  6.613e-02   5.841 6.47e-09 ***
## MSZoningRH           3.368e-01  6.578e-02   5.120 3.50e-07 ***
## MSZoningRL           4.102e-01  5.457e-02   7.516 1.03e-13 ***
## MSZoningRM           3.285e-01  5.151e-02   6.378 2.45e-10 ***
## SaleTypeCon          2.129e-01  1.042e-01   2.044 0.041162 *
## SaleTypeConLD        1.179e-01  5.239e-02   2.251 0.024541 *
## SaleTypeConLI       -4.909e-02  7.369e-02  -0.666 0.505437
## SaleTypeConLw        3.376e-02  6.669e-02   0.506 0.612731
## SaleTypeCWD          1.199e-01  7.375e-02   1.626 0.104213
## SaleTypeNew          2.235e-02  2.803e-02   0.797 0.425423
## SaleTypeOth          1.007e-01  8.455e-02   1.191 0.233970
## SaleTypeWD           3.942e-02  2.244e-02   1.756 0.079239 .
## NeighborhoodBlueste -4.093e-02  1.076e-01  -0.380 0.703699
## NeighborhoodBrDale  -1.305e-01  5.510e-02  -2.367 0.018052 *
## NeighborhoodBrkSide  2.175e-02  4.578e-02   0.475 0.634811
## NeighborhoodClearCr  1.597e-01  4.700e-02   3.399 0.000697 ***
## NeighborhoodCollgCr  5.208e-02  3.682e-02   1.414 0.157464
## NeighborhoodCrawfor  1.647e-01  4.381e-02   3.760 0.000177 ***
## NeighborhoodEdwards -5.756e-02  4.030e-02  -1.428 0.153519
## NeighborhoodGilbert  3.761e-02  3.876e-02   0.970 0.332058
## NeighborhoodIDOTRR  -4.946e-03  5.315e-02  -0.093 0.925878
## NeighborhoodMeadowV -1.079e-01  5.497e-02  -1.963 0.049845 *
## NeighborhoodMitchel  3.887e-03  4.108e-02   0.095 0.924641
## NeighborhoodNAmes    1.669e-02  3.861e-02   0.432 0.665659
## NeighborhoodNoRidge  1.326e-01  4.430e-02   2.992 0.002818 **
## NeighborhoodNPkVill -6.476e-02  5.884e-02  -1.101 0.271303
## NeighborhoodNridgHt  1.308e-01  3.994e-02   3.275 0.001085 **
## NeighborhoodNWAmes   1.373e-02  3.964e-02   0.346 0.729089
## NeighborhoodOldTown -3.524e-02  4.710e-02  -0.748 0.454463
## NeighborhoodSawyer  -1.258e-03  4.060e-02  -0.031 0.975294
## NeighborhoodSawyerW  3.246e-02  3.970e-02   0.818 0.413760
## NeighborhoodSomerst  9.403e-02  4.671e-02   2.013 0.044299 *
## NeighborhoodStoneBr  1.372e-01  4.921e-02   2.787 0.005397 **
## NeighborhoodSWISU    2.175e-02  5.004e-02   0.435 0.663814
## NeighborhoodTimber   9.965e-02  4.241e-02   2.350 0.018937 *
## NeighborhoodVeenker  1.369e-01  5.761e-02   2.377 0.017609 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1393 on 1351 degrees of freedom
## Multiple R-squared:  0.8518, Adjusted R-squared:  0.8466
## F-statistic: 165.2 on 47 and 1351 DF,  p-value: < 2.2e-16
```

Summary function is used in order to check which columns show less correlation to liner model. 3 columns (SaleType,X1stFlrSF,FullBathis) shows less correlation record. When it comes to Neighborhood,Each neighborhood has a different Estimate value from summary.

Remove outliers

```
outlierTest(linear_model)
```

```
##        rstudent unadjusted p-value Bonferroni p
## 1299 -15.879400         3.4370e-52   4.8084e-49
## 524   -8.989204         8.2013e-19   1.1474e-15
## 633   -5.565405         3.1507e-08   4.4078e-05
## 969   -4.881962         1.1755e-06   1.6446e-03
## 1325  -4.509579         7.0590e-06   9.8755e-03
## 463   -4.348741         1.4721e-05   2.0594e-02
```

```
selected_train <-selected_train[-c(463,524,633,969,1299,1325), ]
```

outliers are detected by outlierTest. 6 columns are chosen from the test and these columns is removed to minimize RMSE.

Final model test

```
final_model <- lm(log(SalePrice)~YearBuilt+YearRemodAdd+LotArea+OverallCond+OverallQual + log(GrLivArea
+                      GarageCars  + TotalBsmtSF + MSZoning +
+                    Neighborhood, selected_train)

rmse(final_model, test)
```

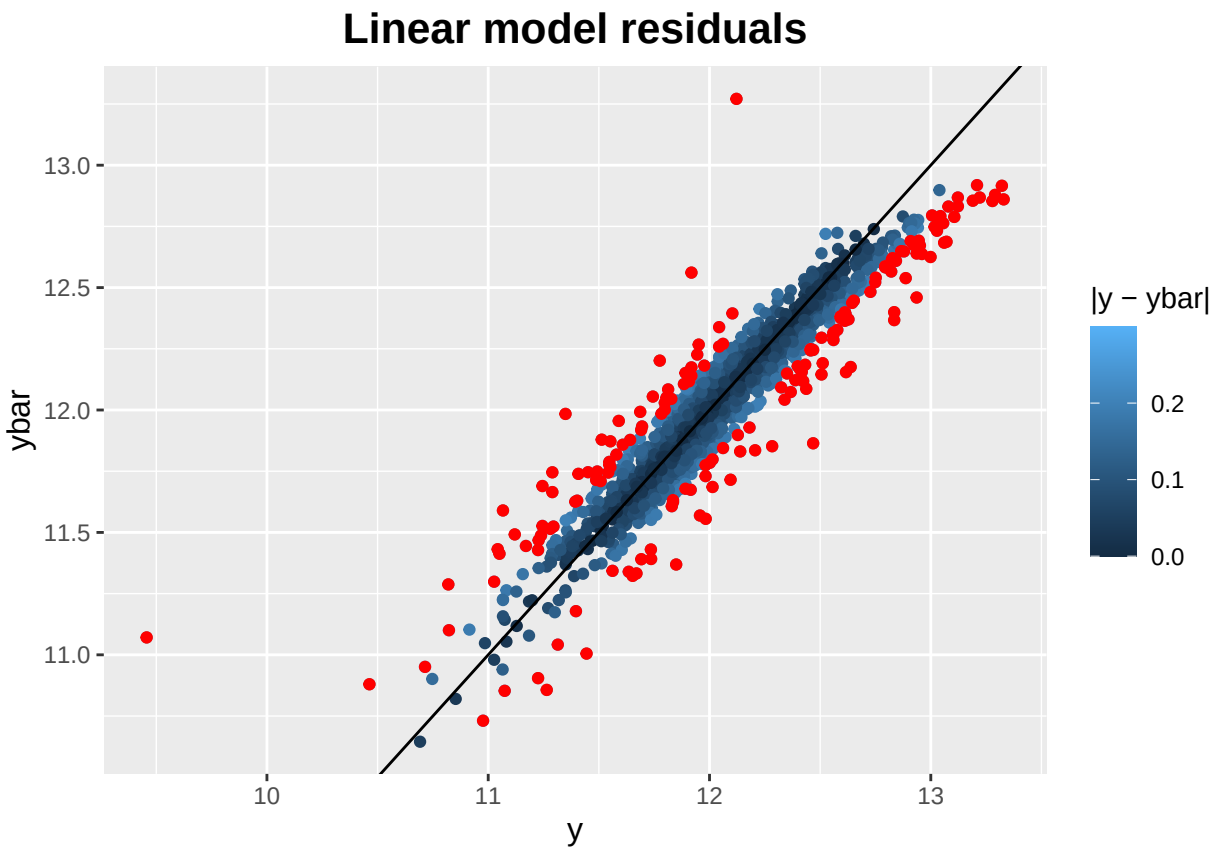
```
## [1] 0.1444709
```

After the first linear model, SaleType,X1stFlrSF,FullBathis, GarageArea are removed and log on GrLivArea variable to fix skewness to improve the model and minimize RMSE.

RMSE First model- 0.148 Final model- 0.144

The RMSE decreased but not significantly from about 0.148 in the first model to 0.144 in the final model. The two models have quite small root mean square error (which is great)on siginificant level at 0.05. The final model has bit less root mean square error so it will be used for further process.

6.EVALUATION




```
## [1] 0.1444709
```

7.RECOMMENDATIONS AND FINAL CONCLUSIONS

In the conclusion, The final model records 0.1443751 of root mean square error. It is very great figure for future prediction.

10 columns is chosen For the final model which are YearBuilt,YearRemodAdd,LotArea,OverallCond,OverallQual,log(GrLivAre GarageCars,TotalBsmtSF,MSZoning,Neighborhood.

The initial model has 14 columns and 4 columns are dropped based on summary(linear model). The outcome is slighly improved which mean RMSE is decreased.

These columns are based on EDA, Futher pre-procssing and modeling parts. These columns show certain pattern to sale price and it is found that it is effective to minimize RMSE. GrLivArea has a skewness so log is applied.

YearRemodAdd, OverallQual, OverallCond,GrLivArea, GarageCars, MSZoning, TotalBsmtSF and Neighborhood are important elements which have an great impacts on sale price so it can tell these columns are key to predict future sale price.

Probably There might be bit more columns which help predict sale price but I do believe this reports check all the major columns which have great impact on sale price. Probably, Analysis could be improved if every outliers is removed in the first stage for every numerical columns.

Students were required to find elements which have correlation to sale price though Pre-processing, EDA, Further processing. However, I think we can use Anova() fucntion to find correlation in out real life quicker.

Actually, It was my first time to organize whole report using R markdown as a report formation. It took a lot of time but it was very meaningful report. Thank you so much Shuan! you are the best teacher in University.

8.REFERENCES

Season in Ames https://www.weather-us.com/en/iowa-usa/ames-climate

Example of project https://www.kaggle.com/c/house-prices-advanced-regression-techniques/notebooks?sortBy=hotness&group=everyone&pageSize=20&competitionId=5407&language=R

Code advice (several pages) https://stackoverflow.com

Code adivce2 (several pages) https://www.r-project.org/

Overall Structure (several pages) http://www.kaggle.com

Understanding of RMSE https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/#:~:text=Root%20Mean%20Square%20Error%20(RMSE)%20is%20the%20standard%20deviat