Exploring the real estate market

Jiaxun Li/Yuchun Wu/Ziqi Zhao/Selena Li

12/16/2022

Introduction

Our group members are really interested in the real estate market, and all members tried their best in order to create a decent report in this area.

House styles are an important aspect of the real estate market, as they can have a significant impact on the value of a property and the appeal it holds for potential buyers. They can also be an important factor in determining the sale price of a property. It can be challenging for people to find the right price to purchase a house, especially if they are not familiar with the local real estate market or if they are unfamiliar with the various factors that can affect the value of a property. Building a model for reference can be a helpful way to provide guidance and support to people who are trying to find the right price to purchase a house. In this analysis, we will explore the various house styles that are popular among different segments of the population, examining factors such as the features and amenities that are most important to people, and the trends and changes that have occurred over time. By gaining a deeper understanding of the value that different house styles hold for people, we can better understand the forces that shape the housing market and make informed decisions about how to navigate it successfully.

About The Data

URL:(https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview)

Data Description

The dataset we use is The Ames Housing Data. The dataset contains 80 variables directly related to the sale of a property and a target value, SalePrice. These 80 variables focus on the types of information that a typical home buyer would like to know about a house. (e.g., when it was built, Heating quality, how many bathrooms it has.) It contains 20 continuous variables related to the various size dimensions of the house. In addition to the basic house size, others include more specific areas such as the basement, porches, etc. The dataset also contains 14 discrete variables, usually the number of kitchens, bedrooms, and bathrooms. Finally, the dataset has 23 nominal variables and 23 ordinal variables. The smallest is STREET, and the largest is NEIGHBORHOOD. Categorical variables usually refer to the environment, garage, material condition (Type of street: Gravel, Paved), and various ratings (Heating Condition: Excellent, Good, Fair). The dataset also contains PIDs, which are identification numbers assigned to each property (like an index) The dataset contains data on houses sold in Ames from 2006 to 2010, with a training dataset length of 1460 and a test dataset length of 1459.

Data Dictionary

- `HouseStyle`:It means the different style of dwelling. Its data type is character.1.5Fin: One and one-half story/SFoyer: Split Foyer/SLvL: Split Level
- `SalePrice`: It means the house sale price. It is quantitative and numeric data.
- `LotArea`:It means lot size in square feet. It is quantitative data. It is numeric and its unit is square feet.
- `YearBuilt`:It means the year house built. It is quantitative and numeric data.
- `Street`: Type of road access to property. Its data type is character.
- `LotShape`: It means the general shape of property. Its data type is character.
- `Utilities`: Type of utilities available. Its data type is character.
- `LotConfig`: It is the lot configuration. Its data type is character
- `OverallQual`: Rates the overall material and finish of the house.
- `OverallCond`: Rates the overall condition of the house. Its data type is character.
- `YearRemodAdd`: It means the remodel date (same as construction date if no remodeling or additions). Its data type is numeric.
- `BsmtQual`: Evaluates the height of the basement. Its data type is character.
- `BsmtCond`:Evaluates the general condition of the basement. Its data type is character.
- `LowQualFinSF`: Low quality finished square feet (all floors)
- `OpenPorchSF`: It means the open porch area in square feet. Its data type is numeric, and its unit is square feet.
- `GrLivArea`: It means the above ground living area square feet. Its data type is numeric.
- `1stFlrSF`: First Floor square feet. Its data type is numeric, and its unit is square feet.
- `2ndFlrSF`: Second floor square feet. Its data type is numeric, and its unit is square feet.

Questions 1

Which housestyle should we recommend people to purchase in this area?

According to my 5 year experience in the US, I tend to choose 1 story house to recommend but we will see what we find in our analysis.

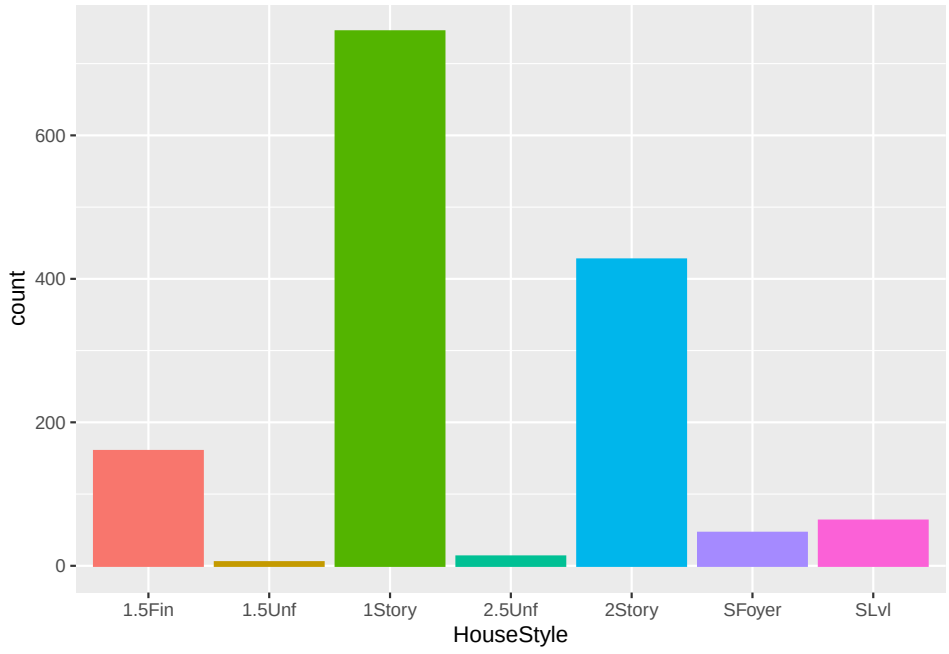
Methodologies

```
kable(sort(table(data.test$HouseStyle),decreasing = TRUE),
           caption="House Styles",
          col.names = c("HouseStyles" , "number"))
```

Table 1: House Styles

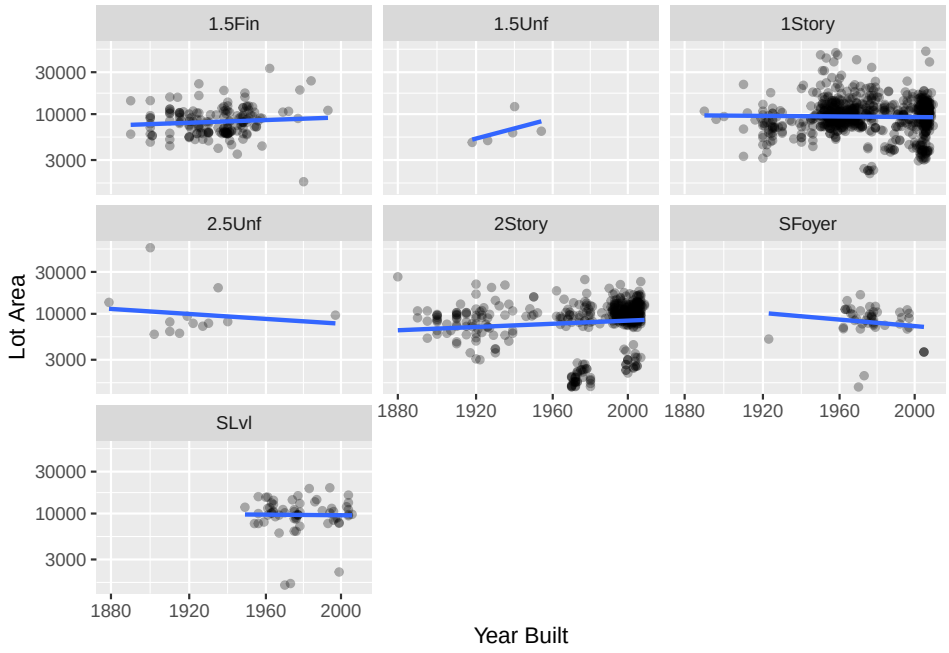
HouseStyles	number
1Story	745
2Story	427
1.5Fin	160
SLvl	63
SFoyer	46
2.5Unf	13
1.5Unf	5

```
ggplot(data = data.test, aes(x = HouseStyle, color = HouseStyle))+
  geom_bar( aes(fill = HouseStyle))+
theme(legend.position = "none")
```

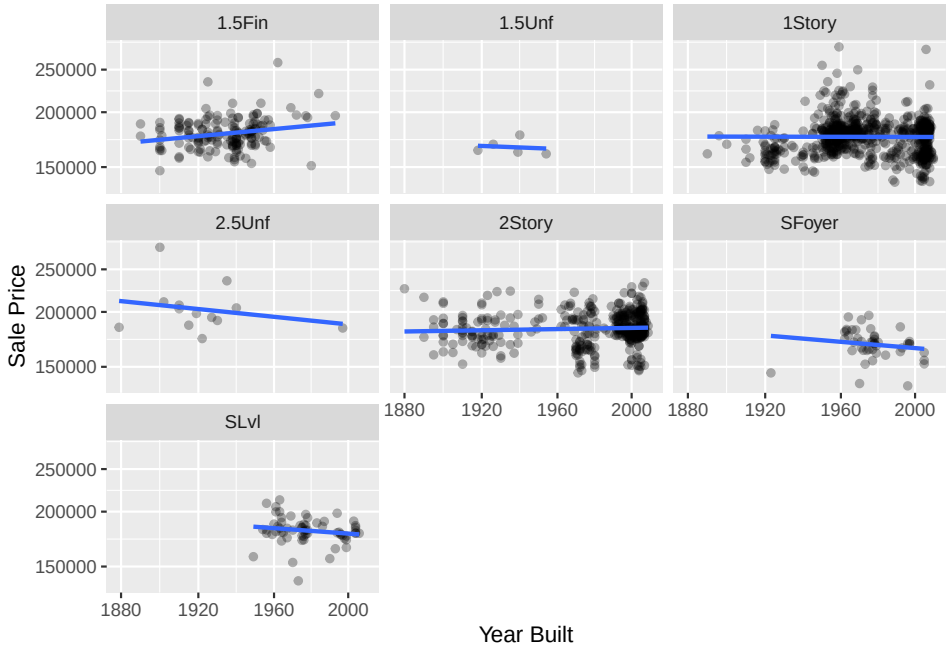


First, I would like to figure out the most popular house style in this area. From the plots, we can see that 1 sotry and 2 stroy house style are most common in this area, which means people may have more choices to pick one desirable house among a large quantity of houses. Also, that means buyers may have hihger bargaining power since the supply of these two styles is higher.

```
ggplot(data = data.test, aes(x = YearBuilt, y =LotArea)) + geom_point(alpha=.3) + geom_smooth(method =
labs(y="Lot Area", x= "Year Built")+ facet_wrap(~HouseStyle)
```




```
ggplot(data = data.test, aes(x = YearBuilt, y =SalePrice)) + geom_point(alpha=.3) + geom_smooth(method =
labs(y="Sale Price", x= "Year Built")+ facet_wrap(~HouseStyle)
```

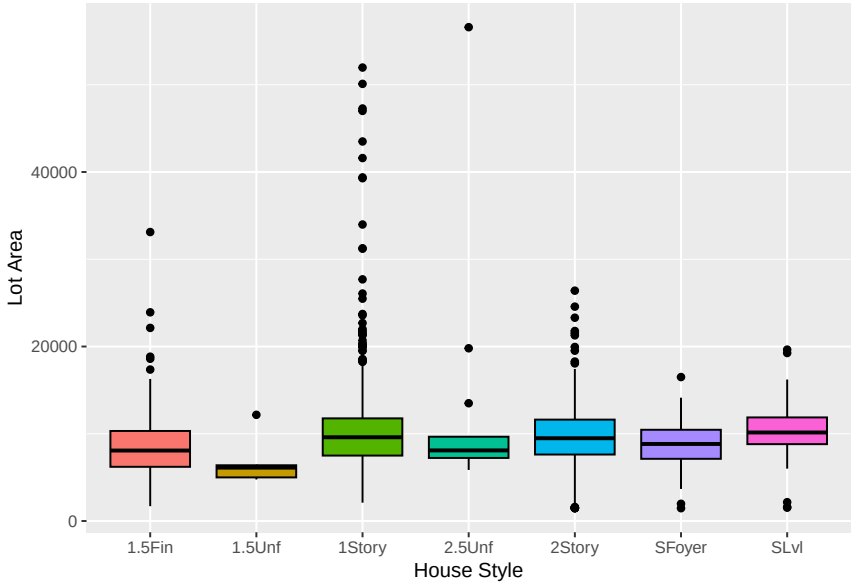


Second, we would like to check the depreciation rate of these two house styles. We want to know their ability to keep values. Before that, we want to eliminate the impact of house sizes on the sale price. From the first graph, we see that, House built in different years are about the same area. Thus, areas would not have large impact on the sale price. Then from the second plot, we can see tha,t for one story and two story houses, there is little difference in sale price for houses built in different years. Thus, these two houses' styles are both good at keeping their value.

```
data.test %>%
group_by(HouseStyle) %>%
summarise(avg = mean(LotArea),median = median(LotArea),
std = sd(LotArea))
```

```
## # A tibble: 7 x 4
##   HouseStyle    avg median    std
##   <chr>       <dbl>  <dbl>  <dbl>
## 1 1.5Fin      8803.   8079   3849.
## 2 1.5Unf      6889.   6120   3035.
## 3 1Story     10345.   9600   5435.
## 4 2.5Unf     12783.   8094  13696.
## 5 2Story      9262.   9487   4140.
## 6 SFoyer      8644.   8832.  2916.
## 7 SLvl       10442.  10147   3480.
```

```
ggplot(data = data.test, aes(x = HouseStyle, y = LotArea, color =HouseStyle))+
  geom_boxplot(color = "black", aes(fill = HouseStyle))+
  theme(legend.position = "none") +
  labs(x="House Style", y = "Lot Area")
```

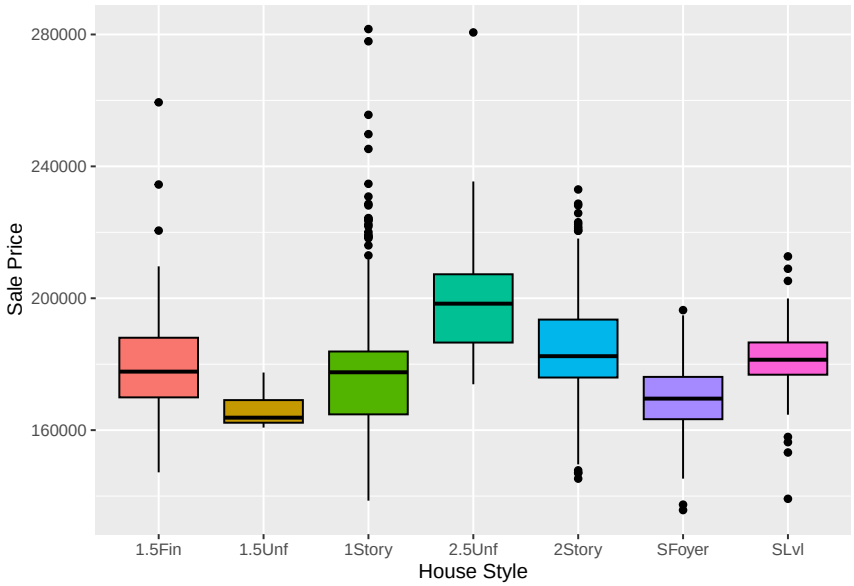


Later on, I would like to see if the living space for these two styles is large enough. From the boxplots, we can see that one-story and two-story house are the top 4 in terms of lot areas. Thus, I believe that the space for them should be enough.

```
data.test %>%
group_by(HouseStyle) %>%
summarise(avg = mean(SalePrice),median = median(SalePrice),
std = sd(SalePrice))
```

```
## # A tibble: 7 x 4
##   HouseStyle     avg  median    std
##   <chr>        <dbl>   <dbl>  <dbl>
## 1 1.5Fin     179307. 177750. 14803.
## 2 1.5Unf     166677. 163785.  6805.
## 3 1Story     176484. 177548. 16593.
## 4 2.5Unf     204184. 198359. 27690.
## 5 2Story     183986. 182427. 15519.
## 6 SFoyer     169055. 169562. 12883.
## 7 SLvl       181484. 181354. 12340.
```

```
ggplot(data = data.test, aes(y= HouseStyle,x= SalePrice, color =HouseStyle))+
  geom_boxplot(color = "black", aes(fill = HouseStyle))+
  coord_flip()+
  theme(legend.position = "none") +
  labs(x="Sale Price", y="House Style")
```



In the last part, which is the most important part, AI would like to compare the price among these seven house styles. We can see that the average sale price for one story house is the second lowest. However, the average sale price for two story house is the second highest.

Results and Conclusion

In conclusion, after we compare the quantities, depreciation rate, lot areas and the sale price of different house styles, we found that the one story house should be the most cost-effective house for the reason that it has the second largest average area with the second lowest average sale price. Also, it has ability to hold its value over times. Besides, It has the largest quantities of houses to choose. Therefore, we recommend people to purchase one-story house.

Question 2

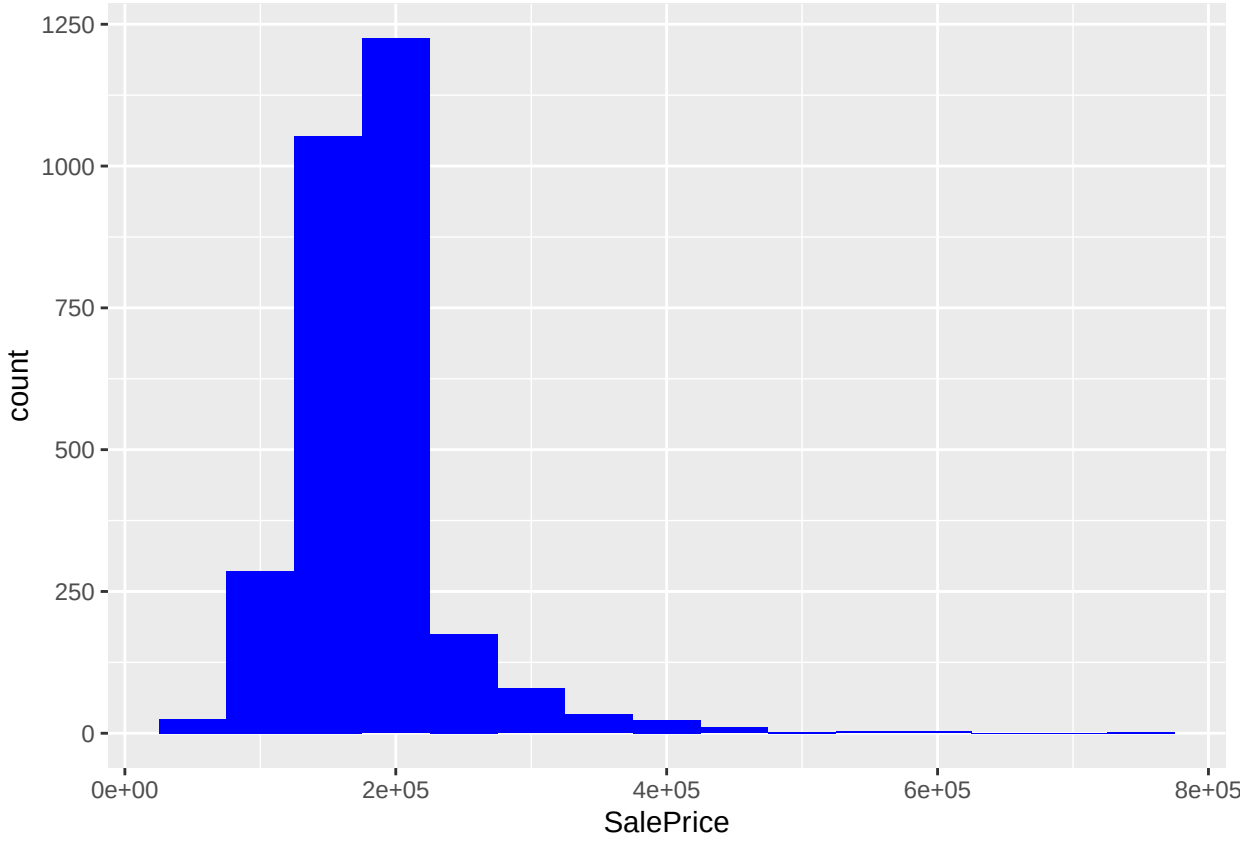
What factors can most effectively estimate the house sale price?

We believe that lotArea,overall condition, and year built may be the most important factors.

Methodologies

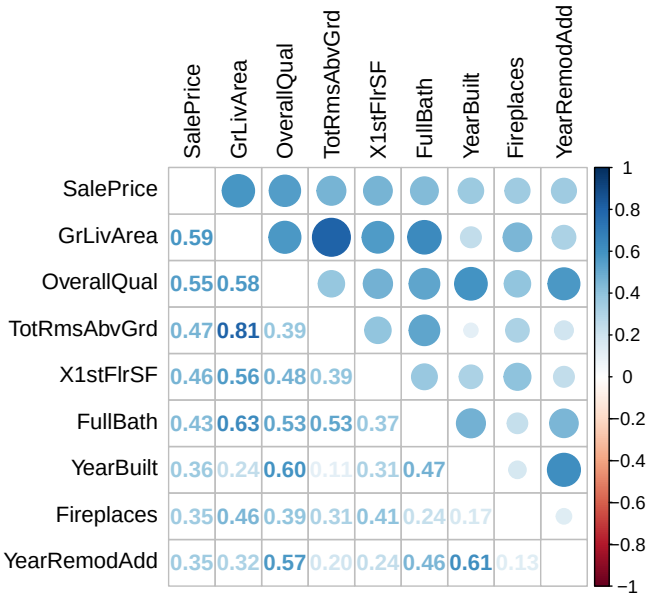
Data Visualization

```
ggplot(data = data_dropna, aes(x = SalePrice)) +
  geom_histogram(fill = "blue", binwidth = 50000)
```



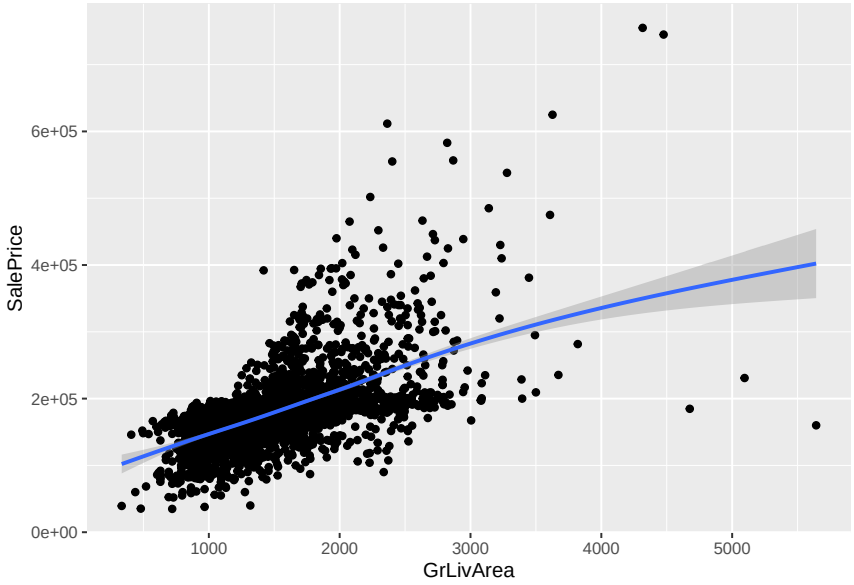
The bar plot here shows that most of the house price in Ames is between 100,000 to 200,000, which is about 2000+ houses' sale price is inside this interval.

```
corr <- cor(numericvars, use = "pairwise.complete.obs")

corr_sort <- as.matrix(sort(corr[,"SalePrice"], decreasing = TRUE))
corr05 <- names(which(apply(corr_sort, 1, function(x) abs(x) > 0.3)))
corr <- corr[corr05, corr05]

corrplot.mixed(corr, tl.col="black", tl.pos = "lt")
```



Here is the Heatmap about the correlation between each numerical variables and house price. The GrLivArea has the highest correlation with house price, which is 0.59. The YearRemodAdd has the lowest correlation with house price, which is 0.35. Also, GrLivArea and TotRmsAbvGrd has strong relationship, which is above 0.8. It is easy to understand that more rooms above ground means more living area above ground. Therefore, the Heatmap is accurate.

```
ggplot(data = data_dropna, aes(x = GrLivArea, y = SalePrice)) +
  geom_point() + stat_smooth()
```



The scatter plot and regression line proof that the correlation between GrLivArea and SalePrice is positive, with a slope about 0.59. However, when the area above group going really high, it has less effect to the house price, according to the graph.

Modeling

Linear Regression

```
lm(SalePrice~., data = num.train) -> linear
summary(linear)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = num.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -301353  -24707   -4290   21076  406063
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.252e+05  1.389e+06   0.306  0.75954
## MSSubClass    -9.843e+01  2.452e+01  -4.014 6.15e-05 ***
## LotArea        9.852e-01  1.252e-01   7.867 5.54e-15 ***
## OverallQual    8.035e+03  1.054e+03   7.625 3.55e-14 ***
## OverallCond    2.633e+03  9.863e+02   2.670  0.00764 **
## YearBuilt      2.943e+02  5.627e+01   5.230 1.85e-07 ***
## YearRemodAdd   1.317e+01  6.505e+01   0.203  0.83952
## X1stFlrSF      4.008e+01  4.377e+00   9.155  < 2e-16 ***
## X2ndFlrSF      3.593e+01  4.756e+00   7.555 6.00e-14 ***
## LowQualFinSF   5.701e+00  1.866e+01   0.306  0.76001
## GrLivArea             NA         NA      NA       NA
## FullBath       1.337e+03  2.624e+03   0.510  0.61040
## HalfBath      -4.784e+03  2.617e+03  -1.828  0.06773 .
## BedroomAbvGr  -2.573e+03  1.580e+03  -1.628  0.10370
## KitchenAbvGr  -1.514e+04  5.116e+03  -2.960  0.00311 **
## TotRmsAbvGrd   5.006e+03  1.182e+03   4.237 2.36e-05 ***
## Fireplaces    -2.725e+02  1.688e+03  -0.161  0.87180
## WoodDeckSF     1.098e+01  7.582e+00   1.448  0.14766
## OpenPorchSF   -8.515e+00  1.436e+01  -0.593  0.55322
## EnclosedPorch  1.828e+01  1.548e+01   1.181  0.23783
## X3SsnPorch     9.405e+00  3.615e+01   0.260  0.79476
## ScreenPorch    3.835e+01  1.609e+01   2.385  0.01718 *
## PoolArea       1.145e+01  2.512e+01   0.456  0.64868
## MiscVal       -4.596e+00  1.471e+00  -3.124  0.00180 **
## MoSold         9.137e+02  3.306e+02   2.764  0.00576 **
## YrSold        -4.953e+02  6.910e+02  -0.717  0.47361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42820 on 2307 degrees of freedom
## Multiple R-squared:  0.4729, Adjusted R-squared:  0.4674
## F-statistic: 86.24 on 24 and 2307 DF,  p-value: < 2.2e-16
```

From the summary of linear regression, there are 7 statistically significant variables: MSSubClass, LotArea, OverallQual, YearBuilt, X1stFlrSF, X2ndFlrSF and TotRmsAbvGrd. Moreover, the p-value for the linear regression is also < 0.05, which means the regression is significant.

```
linear.predict = predict(linear, num.test)
linearr2 <- cor(linear.predict,num.test$SalePrice)^2
cat('R2 for Linear Regression:', linearr2)
```

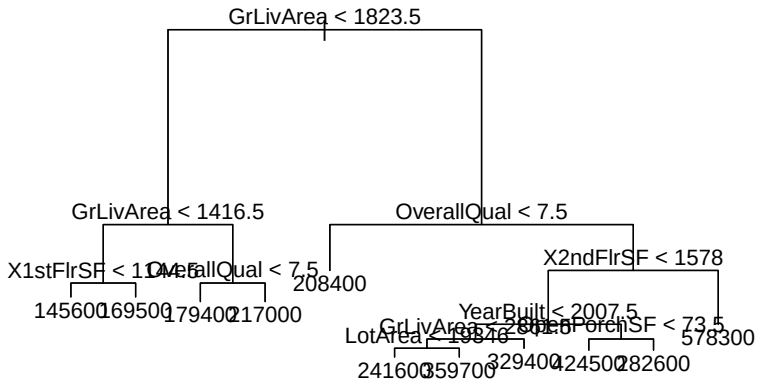
```
## R2 for Linear Regression: 0.4607526
```

Decision Tree

```
tree(SalePrice~., data = num.train) -> tree.numtrain
summary(tree.numtrain)
```

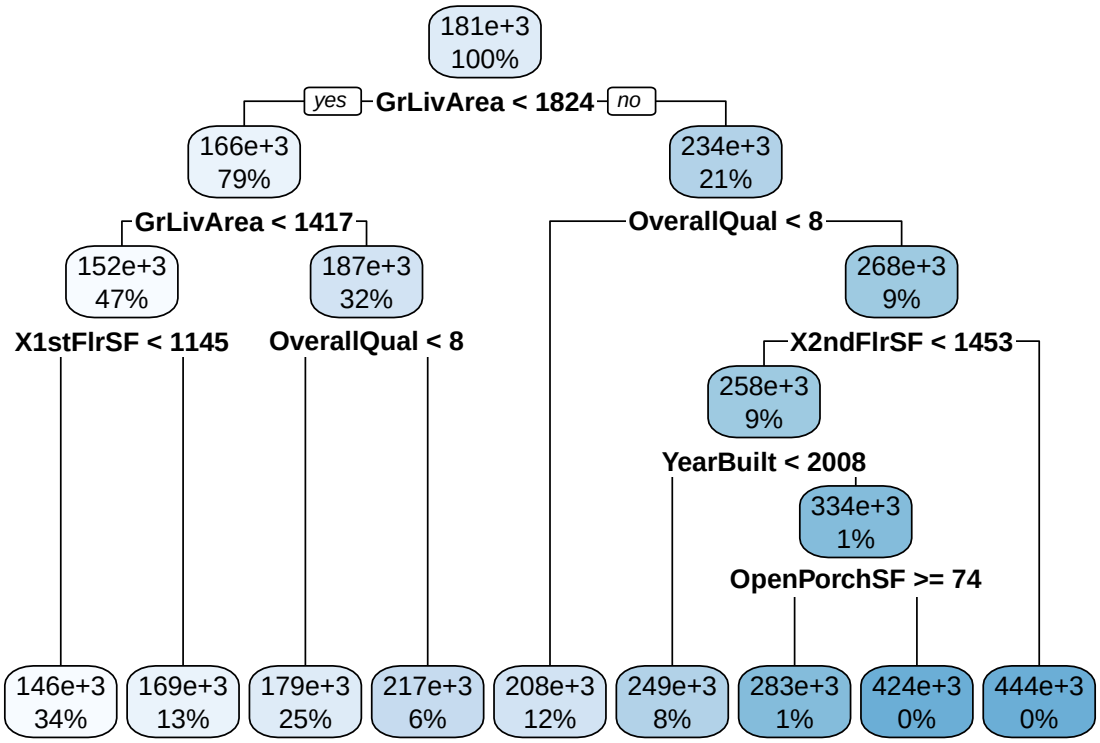
```
##
## Regression tree:
## tree(formula = SalePrice ~ ., data = num.train)
## Variables actually used in tree construction:
## [1] "GrLivArea"   "X1stFlrSF"   "OverallQual" "X2ndFlrSF"   "YearBuilt"
## [6] "LotArea"     "OpenPorchSF"
## Number of terminal nodes:  11
## Residual mean deviance:  1.761e+09 = 4.088e+12 / 2321
## Distribution of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -349700.0  -24450.0     596.5       0.0   20710.0  300300.0
```

```
plot(tree.numtrain)
text(tree.numtrain, pretty = 1)
```



From the plot of decision tree, it filters out 7 important variables. There are two variables that different from linear regression, which are: YearBuilt, OpenPorchSF. The built year from decision tree shows that the house with earlier built year has less sale price, which is conform to the common sense.

```
decision_tree <-
  decision_tree() %>%
  set_engine('rpart') %>%
  set_mode('regression')

#Importance graph
tree_fit <- fit(decision_tree, SalePrice ~., data = num.train)
rpart.plot(tree_fit$fit)
```




```
#Linear Regression with Tree's important variables

lm(SalePrice~OverallQual+GrLivArea+X1stFlrSF+X2ndFlrSF+
     YearBuilt+LotArea+OpenPorchSF, data = num.train) -> lm.tree
summary(lm.tree)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + X1stFlrSF +
##     X2ndFlrSF + YearBuilt + LotArea + OpenPorchSF, data = num.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -323098  -24508   -4069   20930  394830
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.518e+05  7.292e+04  -4.825 1.49e-06 ***
## OverallQual  1.001e+04  9.600e+02  10.432  < 2e-16 ***
## GrLivArea    1.686e+01  1.853e+01   0.910   0.3630
## X1stFlrSF    3.263e+01  1.889e+01   1.727   0.0843 .
## X2ndFlrSF    2.246e+01  1.873e+01   1.199   0.2306
## YearBuilt    1.975e+02  3.836e+01   5.148 2.85e-07 ***
## LotArea      1.107e+00  1.241e-01   8.922  < 2e-16 ***
## OpenPorchSF -7.176e+00  1.433e+01  -0.501   0.6166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43530 on 2324 degrees of freedom
## Multiple R-squared:  0.4515, Adjusted R-squared:  0.4498
## F-statistic: 273.3 on 7 and 2324 DF,  p-value: < 2.2e-16
```

```
tree.predict = predict(lm.tree, num.test)
#Prediction R2
treer2 <- cor(tree.predict,num.test$SalePrice)^2
cat('R2 for Desicion Tree:', treer2)
```

```
## R2 for Desicion Tree: 0.4429462
```

Here is the importance plot of decision tree. From the importance graph, we could see that GrLivArea has highest importance, which is 100%. The second important variable is X1stFlrSF with 47%.

Best Subset

```
regsubsets(SalePrice~OverallQual+GrLivArea+X1stFlrSF+X2ndFlrSF+
             YearBuilt+LotArea+OpenPorchSF,
           data = num.train)->num.train.best
summary(num.train.best)
```

```
## Subset selection object
## Call: regsubsets.formula(SalePrice ~ OverallQual + GrLivArea + X1stFlrSF +
##     X2ndFlrSF + YearBuilt + LotArea + OpenPorchSF, data = num.train)
## 7 Variables  (and intercept)
##             Forced in Forced out
## OverallQual     FALSE      FALSE
## GrLivArea       FALSE      FALSE
## X1stFlrSF       FALSE      FALSE
## X2ndFlrSF       FALSE      FALSE
## YearBuilt       FALSE      FALSE
## LotArea         FALSE      FALSE
## OpenPorchSF     FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          OverallQual GrLivArea X1stFlrSF X2ndFlrSF YearBuilt LotArea
## 1  ( 1 ) " "         "*"       " "       " "       " "       " "
## 2  ( 1 ) "*"         "*"       " "       " "       " "       " "
## 3  ( 1 ) "*"         "*"       " "       " "       " "       "*"
```

```
## 4  ( 1 ) "*"         "*"        " "        " "        "*"        "*"
## 5  ( 1 ) "*"         " "        "*"        "*"        "*"        "*"
## 6  ( 1 ) "*"         "*"        "*"        "*"        "*"        "*"
## 7  ( 1 ) "*"         "*"        "*"        "*"        "*"        "*"
##          OpenPorchSF
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) "*"
```

```
which.max(num.train.best$adjr2)
```

```
## integer(0)
```

```
coef(num.train.best, 3)
```

```
##  (Intercept)  OverallQual    GrLivArea      LotArea
## 25142.991845 13558.790804    39.992627     1.222029
```

We first use the best subset regression because we want to improve the out-of-sample accuracy of the regression model by eliminating the unnecessary predictors. From the output, we can see that the best subset selection with three predictors is OverallQulaity, GroundLivingArea, and LotArea. By comparison, the OpenPorchSF seems less important.

Forward

```
#Forward
regsubsets(SalePrice ~ OverallQual+GrLivArea+X1stFlrSF+X2ndFlrSF+
             YearBuilt+LotArea+OpenPorchSF,
           data = num.train,method = "forward")->num.train.fwd
summary(num.train.fwd)
```

```
## Subset selection object
## Call: regsubsets.formula(SalePrice ~ OverallQual + GrLivArea + X1stFlrSF +
##     X2ndFlrSF + YearBuilt + LotArea + OpenPorchSF, data = num.train,
##     method = "forward")
## 7 Variables  (and intercept)
##             Forced in Forced out
## OverallQual     FALSE      FALSE
## GrLivArea       FALSE      FALSE
## X1stFlrSF       FALSE      FALSE
## X2ndFlrSF       FALSE      FALSE
## YearBuilt       FALSE      FALSE
## LotArea         FALSE      FALSE
## OpenPorchSF     FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: forward
```

```
##          OverallQual GrLivArea X1stFlrSF X2ndFlrSF YearBuilt LotArea
## 1  ( 1 ) " "         "*"       " "       " "       " "       " "
## 2  ( 1 ) "*"         "*"       " "       " "       " "       " "
## 3  ( 1 ) "*"         "*"       " "       " "       " "       "*"
## 4  ( 1 ) "*"         "*"       " "       " "       "*"       "*"
## 5  ( 1 ) "*"         "*"       "*"       " "       "*"       "*"
## 6  ( 1 ) "*"         "*"       "*"       "*"       "*"       "*"
## 7  ( 1 ) "*"         "*"       "*"       "*"       "*"       "*"
##          OpenPorchSF
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) "*"
```

```
coef(num.train.fwd, 3)
```

```
##  (Intercept)  OverallQual    GrLivArea      LotArea
## 25142.991845 13558.790804    39.992627     1.222029
```

We then use the forward stepwise regression and add variables that improve the model most, one at a time to meet the criteria. The seven predictors of forward stepwise subset selection are OverallQual, GrLivArea, X1stFlrSF, X2ndFlrSF, YearBuilt, LotArea, and OpenPorchSF. Thus, the forward stepwise subset selection we use is the same as the best subset selection.

```
#Linear Regression with Subset's important variables
lm(SalePrice~OverallQual+GrLivArea+LotArea, data = num.train) -> lm.subset
summary(lm.subset)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + LotArea, data = num.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -304441  -24841   -3135   20255  395345
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25142.992   4079.904   6.163 8.41e-10 ***
## OverallQual 13558.791    792.477  17.109  < 2e-16 ***
## GrLivArea      39.993      2.282  17.525  < 2e-16 ***
## LotArea         1.222      0.121  10.098  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43920 on 2328 degrees of freedom
## Multiple R-squared:  0.4404, Adjusted R-squared:  0.4397
## F-statistic: 610.8 on 3 and 2328 DF,  p-value: < 2.2e-16
```

```r
subset.predict = predict(lm.subset, num.test)

#Prediction R2
subsetr2 <- cor(subset.predict,num.test$SalePrice)^2
cat('R2 for Linear Regression:', linearr2,
    'R2 for Desicion Tree:', treer2,
    'R2 for Best Subset:', subsetr2)
```

```
## R2 for Linear Regression: 0.4607526 R2 for Desicion Tree: 0.4429462 R2 for Best Subset: 0.4238167
```

Finally, we made a Linear Regression with Subset's important variables. As shown in the output, the proportion of variability explained by the model is 44%.

Results and Conclusion

According to the three models, the proportion of variability explained by the linear regression, decision tree, and subset selection, respectively, is 46%, 44%, and 42%. According to the above analysis, in the future, if people would like to purchase a house, overall quality, above ground living area, First floor area, second floor area, year bulit, and lot area are the top factors that they should consider. And if people have trouble finding the right price to purchase the house, they can use our regression model to estimate the price.

Sale Price = (4.252e+05) + (9.852e - 01)*Lot Area + (8.035e+03)*Overall Quality + (2.943e+02) * Year Built + (4.008e+01)* 1st Floor Squeard Feet + (3.593e+01)* 2st Floor Squared Feet