House Price Prediction Project

Columbia University

Echo Liu hl3250

1. Define the question or problem that you are trying to solve with your analysis, and explain why it is important or useful, and to whom. Your introduction and background section, should include: (10pts)
a. Some background research (citing at least 1-2 published sources such as journal articles or a textbook) on work done and the types of methods used in the past for this or related problems in that domain;

How to determine home value has been a common problem for every household. Knowing how to calculate your home's value with the help of online valuation tools and trained professionals better prepares you to buy, sell, refinance or even negotiate lower property taxes. According to Nerdwallet website (https://www.nerdwallet.com/blog/mortgages/how-to-determine-home-value/) , there are some methods that have been used to predict house price: online valuation tools including automated valuation model with a confidence score; comparative market analysis by local real estate agent; FHFA House Price Index Calculator by Federal Housing Financing Agency; and a professional appraiser to evaluate the market, the property and comparable properties. It is important for households to know your house value, which allows you to evaluate what you can afford, determine whether a listing is priced appropriately and decide how to price your own home.

b. One or more hypotheses about what you expect to discover (this could be very general or more specific (ie. a specific prediction of which features will be able to predict your outcome);

The objective of this project is to utilize data visualization, feature selection and feature engineering, and machine learning models to predict house prices. Additionally, it is aimed to minimize the difference between predicted and actual rating (RMSE /MSE)
The house price prediction dataset contains 79 explanatory variables describing the majority aspects of residential homes in Ames, Iowa, and 1460 observations in training data while the testing data consists of 1459 observations.

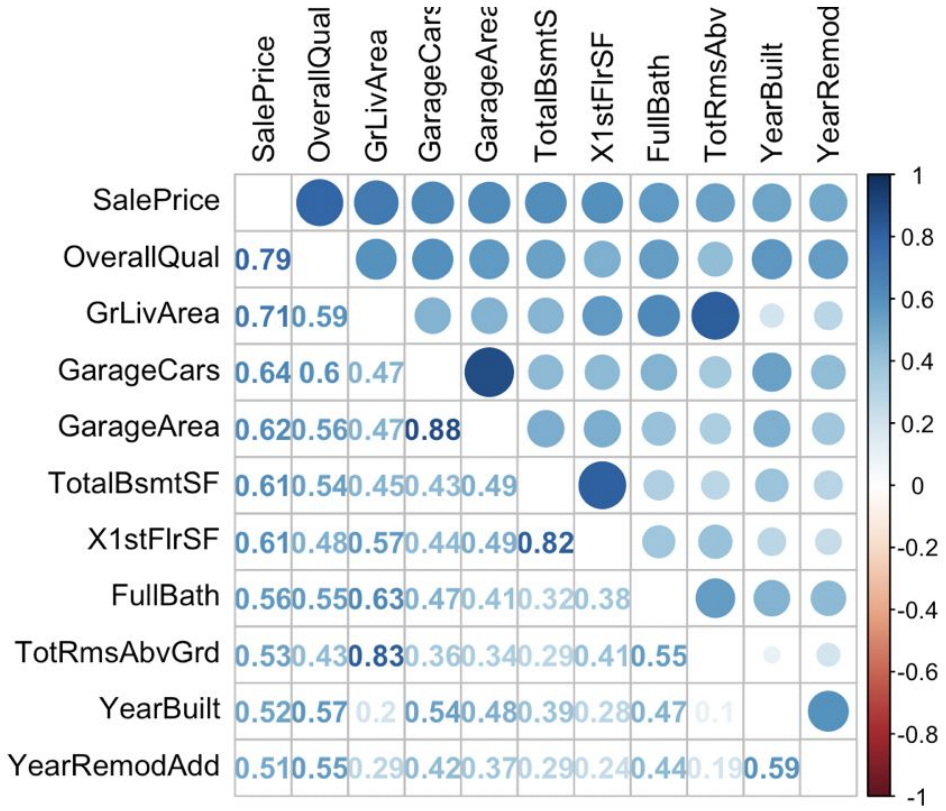
Hypotheses: What variables are highly relevant to determine the sales price? To be more specific, whether the following features including LotArea, Neighborhood, OverallQual, Full Bath, YearBuilt, TotalBsmtSF, CentralAir, GrLivArea, GarageArea, will be able to predict the outcome, the housing price.

c. Consideration of possible social or ethical implications or risks involved in this analysis

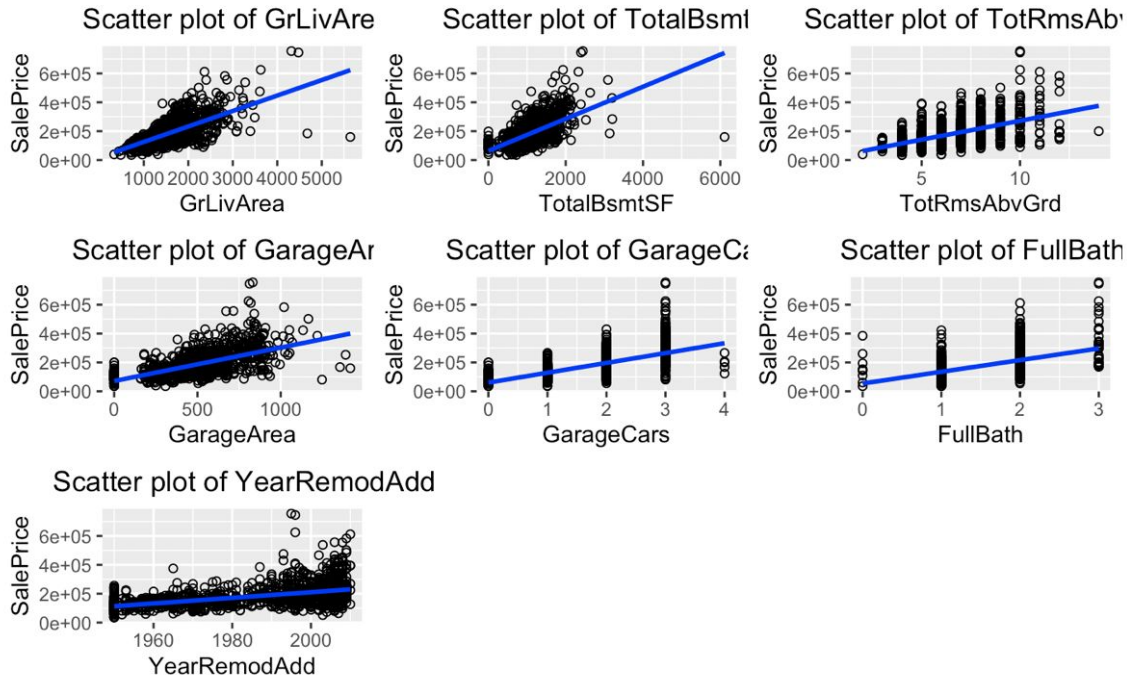
Housing prices are an important reflection of the economy, and housing price ranges are of great interest to both buyers and sellers. Utilizing machine learning models to predict the house price would be very valuable for real state agents who could make use of the information provided on a daily basis.

2. Explore and summarize the dataset using plots and summary statistics. (10 pts)

The training dataset has 1460 observations and 81 variables, and 1459 observations and 80 variables are in the test dataset. In order to get a better understanding and comprehensiveness of the dataset, I decided to first see the top 10 variables that have a high correlation with the SalePrice.



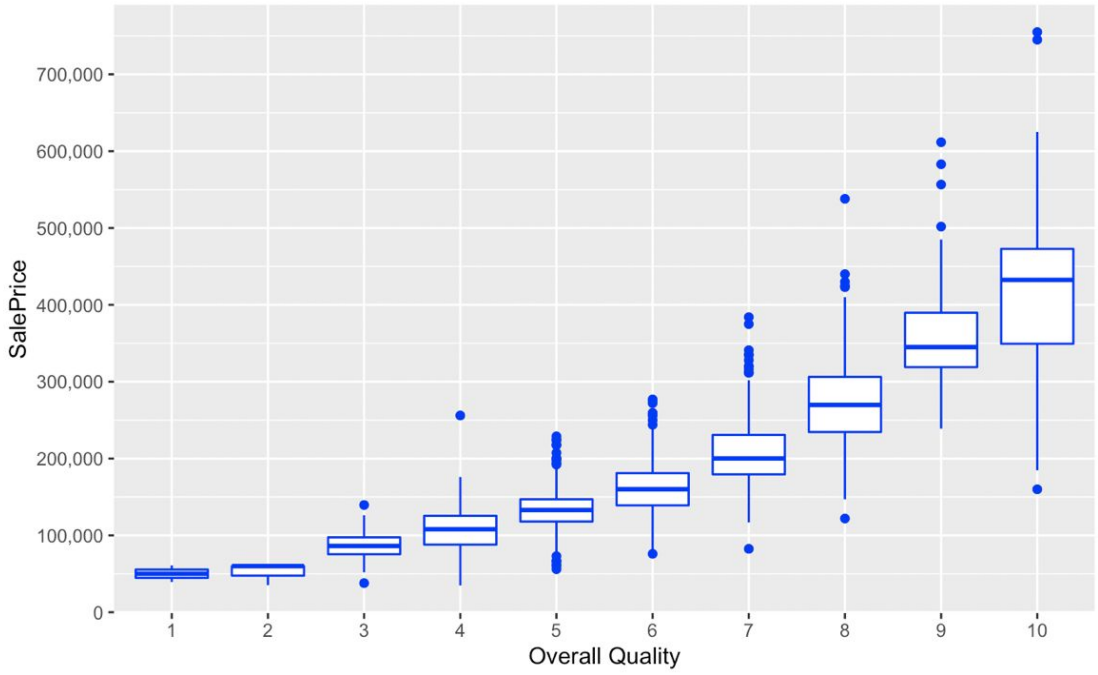
As above, there are 10 out of 38 variables with a positive correlation of at least 0.5 with SalePrice. It also shows the multicollinearity issue which we will discuss in the limitation part. If we think about these variables, we can conclude that they give almost the same information so we only need to choose one of them as key variables.



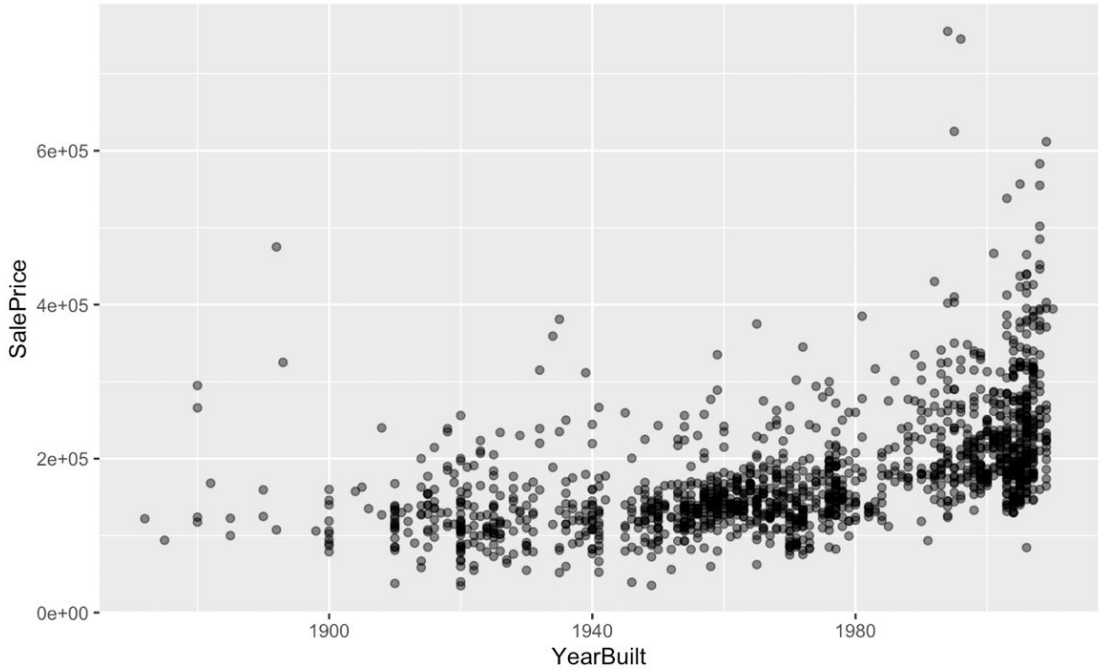
The scatter plot below confirms my opinion.GrLivArea, TotalBsmtSF, TotRmsAbvGrd, and GarageArea are positively correlated with SalePrice, which means that as one variable increases, the other also increases. In the case of 'TotalBsmtSF', we can see that the slope of the linear relationship is particularly high, and it makes sense that big houses are generally more expensive.

One of the figures we may find interesting is the one between 'TotalBsmtSF' and 'GrLiveArea'. In this figure, we can see the dots drawing a linear line, which almost acts like a border. It totally makes sense that the majority of the dots stay below that line. Basement areas can be equal to the above-ground living area, but it is not expected a basement area bigger than the above-ground living area.

In addition, I take an extra look at OverallQual since it has the highest correlation rate with SalePrice. Below are the plots.



Obviously, there is a positive correlation between OverallQual and SalePrice.
I also plot the YearBuilt variable and there is also a positive correlation relationship with SalePrice based on the plot below.



To summarize, I decide to use 8 key variables to predict the models: OverallQual, GrLivArea, TotalBsmtSF, TotRmsAbvGrd, GarageArea, FullBath, YearBuilt, YearRemodAdd.
The next step is checking missing data. Many real-world data-sets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders. Training a model with a data-set that has a lot of missing values can drastically impact the machine learning model's quality.

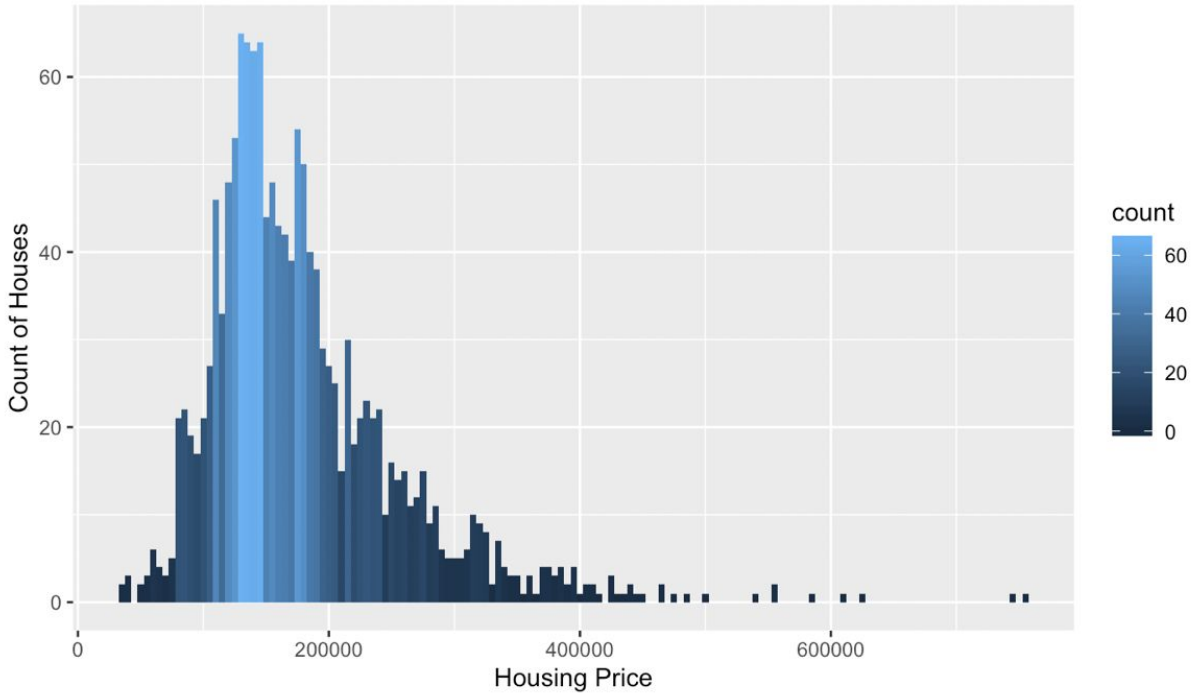
```
      PoolQC  MiscFeature        Alley        Fence  FireplaceQu  LotFrontage   GarageType
GarageYrBlt GarageFinish   GarageQual   GarageCond BsmtExposure
        1453         1406         1369         1179          690          259           81
81           81           81           81           38
BsmtFinType2     BsmtQual     BsmtCond BsmtFinType1   MasVnrType   MasVnrArea   Electrical
          38           37           37           37            8            8            1
```

We'll consider that when more than 15% of the data is missing, we should delete the corresponding variable and pretend it never existed. This means that we will not try any trick to fill the missing data in these cases. According to this, none of these variables seems to be very important since most of them are not aspects in which we think about when buying a house (maybe that's the reason why data is missing?). Moreover, looking closer at the variables, we could say that variables like 'PoolQC', 'MiscFeature' and 'FireplaceQu' are strong candidates for outliers, so we'll be happy to delete them. In addition, there are no NA values in my selected variables, I skipped the part of handling Null Values. However, it brings some major issue that I will discuss in the limitation part.

3. Develop a model to predict the outcome variable(s). Demonstrate how you chose what kind of model to employ, and which variables you chose to include in the model. (20 pts) Your analysis must include:
a. careful analysis of model performance with a train/test/ validation split.
b. evaluation of at least three different methods that were used in the course, using rigorous methods to compare performance between models.

SalePrice is our target variable and also the dependent variable for prediction. Before building models to predict the outcome, SalePrice, there is a need to analyze the outcome itself.

Figure 1 Histogram of SalePrice



As we can see, the sale prices are right-skewed, this was expected because few people can afford highly expensive houses.

In data normalization, I split the training, test, and validation data.

```r
```{r}
set.seed(5678)
data2 = select(train, OverallQual, GrLivArea, TotalBsmtSF, TotRmsAbvGrd, GarageArea, FullBath, YearBuilt,YearRemodAdd,SalePrice)
# Normalize the data
data2[,-9] = as.data.frame(scale(data2[,-9]))
trainid <- createDataPartition(data2$SalePrice, p = 0.8,list = FALSE,times = 1)
train2 <- data2[trainid,]
other = data2[-trainid,]

test_id = createDataPartition(other$SalePrice, p = 0.5,list = FALSE,times = 1)
test2 <- other[test_id,]
valid <- other[-test_id,]

head(train2)
#as.data.frame(table(train$Utilities))
```
```

Once the data is processed we will now proceed further to make our machine learning model. Next step would be the feature selection detailed in section c, and building models, including linear regression model, random forest model, and neural network model.

Linear Regression Model:
As our target variable is continuous we will fit a regression model to the dataset. The aim of this model is to minimize the sum of the squared residuals. Here I select 8

variables to fit into this model: OverallQual , GrLivArea , TotalBsmtSF , TotRmsAbvGrd , GarageArea , FullBath , YearBuilt , YearRemodAdd. I first find outliers and remove them in the dataset. Then I divide datasets into three parts -- training, test, and validation, to prepare for prediction later. Then I ran the model to calculate the RMSE. The RMSE for test data is the best result comparing to the other two.

```
longer object length is not a multiple of shorter object length[1] -1.027719
longer object length is not a multiple of shorter object length[1] 102974.6
```

```{r}
# RMSE of test data
pred = predict(linear_model2, newdata = test2)
sse = sum((pred - test2$SalePrice)^2)
sst = sum((mean(test2$SalePrice)-test2$SalePrice)^2)
model_r2 = 1 - sse/sst; model_r2
rmse = sqrt(mean((pred-test2$SalePrice)^2)); rmse

# RMSE of validation data
pred = predict(linear_model2, newdata = valid)
sse = sum((pred - valid$SalePrice)^2)
sst = sum((mean(valid$SalePrice)-valid$SalePrice)^2)
model_r2 = 1 - sse/sst; model_r2
rmse = sqrt(mean((pred-valid$SalePrice)^2)); rmse
```

```
[1] -370.2693
[1] 1518093
[1] -375.0945
[1] 1515402
```

Random Forest Model:
The RMSE for the model are as follows:

3.2 Random Forest Model

```{r}
# Predict the house price using random forest
set.seed(100)
forest = randomForest(SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF +
TotRmsAbvGrd + GarageArea + FullBath + YearBuilt + YearRemodAdd,train2)


# RMSE of training data
pred = predict(forest)
sse = sum((pred - train2$SalePrice)^2)
sst = sum((mean(train2$SalePrice)-train2$SalePrice)^2)
model_r2 = 1 - sse/sst; model_r2
rmse = sqrt(mean((pred-train2$SalePrice)^2)); rmse

# RMSE of test data
pred = predict(forest, newdata = test2)
sse = sum((pred - test2$SalePrice)^2)
sst = sum((mean(test2$SalePrice)-test2$SalePrice)^2)
model_r2 = 1 - sse/sst; model_r2
rmse = sqrt(mean((pred-test2$SalePrice)^2)); rmse

# RMSE of validation data
pred = predict(forest, newdata = valid)
sse = sum((pred - valid$SalePrice)^2)
sst = sum((mean(valid$SalePrice)-valid$SalePrice)^2)
model_r2 = 1 - sse/sst; model_r2
rmse = sqrt(mean((pred-valid$SalePrice)^2)); rmse
```

```
[1] 0.8230377
[1] 33504.05
[1] 0.8488626
[1] 30629.5
[1] 0.9010785
[1] 24576.77
```

As we can see, validation data performs the best with Random Forest Model.

Neural Network Model:
Variable used: OverallQual , GrLivArea , TotalBsmtSF , TotRmsAbvGrd , GarageArea , FullBath , YearBuilt , YearRemodAdd.

3.3 Neural Network Model

```{r}
# Predict the house price with Neural Network
network = nnet(SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF +
    TotRmsAbvGrd + GarageArea + FullBath + YearBuilt +
    YearRemodAdd, train2, size = 8, linout = TRUE)



# RMSE of training data
pred = predict(network)
sse = sum((pred - train2$SalePrice)^2)
sst = sum((mean(train2$SalePrice)-train2$SalePrice)^2)
model_r2 = 1 - sse/sst; model_r2
rmse = sqrt(mean((pred-train2$SalePrice)^2)); rmse

# RMSE of test data
pred = predict(network, newdata = test2)
sse = sum((pred - test2$SalePrice)^2)
sst = sum((mean(test2$SalePrice)-test2$SalePrice)^2)
model_r2 = 1 - sse/sst; model_r2
rmse = sqrt(mean((pred-test2$SalePrice)^2)); rmse

# RMSE of validation data
pred = predict(network, newdata = valid)
sse = sum((pred - valid$SalePrice)^2)
sst = sum((mean(valid$SalePrice)-valid$SalePrice)^2)
model_r2 = 1 - sse/sst; model_r2
rmse = sqrt(mean((pred-valid$SalePrice)^2)); rmse
```

```
 # weights:  81
 initial  value 45819223631103.796875
 iter  10 value 3130703509268.227539
 final  value 2958564957968.824707
 converged
 [1] 0.6010186
 [1] 50307.57
 [1] 0.590969
 [1] 50388.56
 [1] 0.6256822
 [1] 47807.9
```

For Neural Network Model, validation data outperform than other data with RMSE .

c. consideration of feature selection, transformations, or feature engineering steps

As we can see in the plot above, the distribution of 'SalePrice' is right-skewed which is positive. We would like to get the skewness factor as close to zero as possible. This can be accomplished by either removing outliers or transforming the variable. Removing outliers may be tricky as expertise in real estate is needed to assess whether outliers should be removed or not. Applying transformations is typically a safer option if it can deliver the desired outcome. In the case of positive skewness, log transformation does the trick.

Figure 2 Histogram of log SalePrice

In addition, I use forward stepwise feature selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the tolerance level. If a nonsignificant variable is found, it is removed from the model.

```
Step:  AIC=24716.95
SalePrice ~ OverallQual + GrLivArea + GarageArea + TotalBsmtSF +
    YearBuilt + YearRemodAdd + FullBath

               Df  Sum of Sq           RSS   AIC
<none>                       1755693763084 24717
+ TotRmsAbvGrd  1 1901561951 1753792201133 24718
```

4. What does your model tell you about the data, and about the problem that you are considering? (10 pts)

All three models result in very high RMSE score. Large RMSE indicates that there is a need to include more variables instead of only 8 out of 80 to reduce the RMSE. On the other hand, it indicates that all variables I select are essential for house price prediction. The overall material and finish of the house, original construction date, remodel date, total square feet of basement area, above grade living area square feet, full bathrooms above grade, total rooms above grade and size of a garage in square feet are the most important features when purchasing a house. It solves the problem that people get lost when considering all aspects of a house, which provides a clear and statistical proven range on aspects that really matters for house price.
For the linear regression model, the Adjusted R-squared value is 0.7983, meaning adjusts the R-squared based on the number of independent variables in the model.
For random forest model and neural network, validation data has the smallest RMSE, indicating there is no overfitting problem.

5. Discuss the limitations that you have discovered in the data and in your model. Are there potential sampling issues? Sources of bias? (10 pts)

Variable Selection:
The number of variables I selected is small, although they are highly correlated with the outcome, it is not comprehensive and fully representative and there are lots more correlated variables I failed to analyze.
Handling missing value:
For time-saving and convenience consideration, I got rid of the observations that have missing data. However, I risk losing data points with valuable information. There are some important and correlated variables with missing values such as BsmtQual,

BsmtCond, and GarageQual. The results would be more accurate if I filled in median numbers in NA values or impute them by proceeding sequentially through features with missing values.

Multicollinearity:

The corrplot above shows the multicollinearity issue. For instance, the correlation between GarageCars and GarageArea is very high at 0.89, and both have similar correlations with SalePrice. Same with TotRmsAbvGrd and GrLivArea, TotalBsmtSF and X1stFlrSF. These cases show how significant the correlation is between these variables, this correlation is so strong that it can indicate a situation of multicollinearity.

Outliners:

Outliers is also something that we should be aware of, because outliers can markedly affect our models and can be a valuable source of information, providing us insights about specific behaviours. Outliers is a complex subject and it deserves more attention. However, due to the limit of time and energy, I failed to analyze the outliers through the standard deviation of SalePrice. If time is generous, I would like to do som univariate analysis and bivariate analysis.

Variable transformation:

There are some numerical variables that are really categorical such as OverallCond, YrSold and MoSold. I can also do some label encoding to some categorical variable that may contain information in their ordering sets, such as ExterQual, PoolQC, and CentralAir.

6. What additional data would you like to have? Propose some additional data that you believe would be useful to collect or experiments you could run to validate your observations? (5 pts)

Some house description in text format would be helpful for this dataset by sentimental analysis. Description from different groups of people would be more objective and diverse, sellers and agents may over compliment the house whereas to the contrary, buyers may underestimate the house to get a better deal. The sentimental analysis would help each side of representatives to understand the social sentiment of their house, while monitoring online conversations.

7. Summarize your results and conclusions. (10 pts)

a. How do your findings compare to what you hypothesized?

b. How do your findings relate to earlier work?

c. Are there any ethical issues or risks suggested by your results? Are there any surprises?

d. Why are your results interesting and important?

Based on the results of all my models, the findings stand together with my original hypothesis that besides the most important layout of a house, the availability of garage area, building year, total square feet of basement and number of full bathrooms are essential when predicting the house price.

Relate to earlier research work on real estate, we can see that as an important property type, the price of houses depends on a lot of different variables, including the layout, condition, function, location of the house, as well as some macroeconomic factors. Thus, when predicting the prices, all these variables need to be taken into consideration, or, the result will be too restricted and deviate from the true market value.

Analyzing the neighbourhood into house prediction could be a bit tricky. The education and social level of neighbourhood have an important impact that may raise some social issues. If the education level of surroundings is relatively low, the house price would more likely to drop, which seems unfair. I was surprised that FullBath variable is highly correlated with house price. Personally, it would be a small portation to consider when I purchase the house, but not in the statistical findings.

As we state at the beginning of the project, knowing the correlated variables that would heavily affect the house price would provide a more accurate price prediction. It provides lots of useful information when we consider purchasing a house on a later day, such as garage areas, basement quality, full bath and building year, preventing us from agents and agencies that may hide the real truth and exaggerate unimportantly features that shake our minds.