Predictive Analysis and Techniques on Iowa Housing Data

Abdullah Arshad
500503498
Ceni Babaoglu

Introduction:

Real estate is constantly being bought and sold all around the world everyday at fluctuating prices and Canadian metropolitans face some of the priciest housing options of them all. What I want to build is a system or solution for homeowners to correctly predict the price of the home in their ideal location. Using the Iowa Housing dataset, I will be investigating the many characteristics provided for about 1500 homes and will investigate how important each attribute is in the prediction of house prices. This will require me to evaluate the importance of each attribute and rank them in importance. I will need to filter and clean my data and decide what will and should be used for the final simulation. I will also offer different predictive models and compare to see which one is better suited for our needs. Going past their project and knowing that each housing community has its own differences, my hope is that with what I can create a working dataset for my region; Toronto. With what I present in my report, I hope that any other individual who wants to create a housing dataset for their region could create a concise one, to meet their own goals. The Iowa dataset consists of 79 explanatory variables that describe various criteria's that would influence a house's sale price. I will be completing my project in R, using RStudio's.

Literature Review:

Leading up this project, I tasked myself to read literature about the project target, about Data analytic techniques, and comprehensive methods of incorporating R in my analysis. Here is a list of different literature's that I read while preparing this project.

Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

This is publication from the author of the dataset, Dean De Cock. Although this article did not provide too much insight from my work, it was an excellent introductory guide for the dataset I had chosen. The author goes well to explain how this dataset differs from its more well-known predecessor, Boston Housing, and why he chose to create a newer more complex dataset. The objective of the author is to create a regression problem for his students to solve, however I will be using a different approach, but regardless the tips he used to do the EDA will be valuable in my initial analysis.

Analyzing Spatial Heterogeneity of Housing Prices Using Large Datasets

This publication for an American team discusses the limitations of traditional Spatial econometric methods in the hedonic model and its standard approach in understanding the determinants of housing value, assessing housing values and calculating property tax. The biggest limitation is due to the vast size of housing datasets and their observations. This group of scientists talk about the shortcoming for classical common algorithms for housing classifications, such as; Decision Trees, decision rules, K-Means clustering and KNN classification. The paper seeks to crate an efficient and accurate method for data-driven housing-submarket classification. This paper is interesting because they use their own city and develop their own model, as well as their own city's data. They put forward their innovative hybrid classification system, and I appreciate the process they came to make it.

Robust Variable Selection with Exponential Squared Loss

This publication from a team of Chinese data science students is about different variable selection methods and evaluating to find the most robust one. This a journal is very mathematically intensive and weighs different variable selection techniques. In this, they use the Boston Dataset, and the team subjects the data to three different *Least Absolute shrinkage and selection Operator* (LASSO) techniques, as well as the classic MM-estimator and *Ordinary least squares* (OLS) methods.

Model averaging based on rank

This publication is from another Chinese team in Beijing, where they discuss statistical model selection. The compare the traditional models (Akaike Information Criterion, Bayesian Information Criterion) with more focused information criterions and modeling averages. They subjected the Boston Dataset attribute to Full, AIC, BIC, FIC, and S-FIC methods to estimate the coefficient of each. This team also references the previous publication, Robust Variable Selection with Exponential Squared Loss, and uses criticizes some of the point they make regarding the robustness of the LASSO techniques mentioned. They also say that perhaps the model they choose in that article might be better at selecting models with overall good properties, but not for estimating specific parameter under focus. After going through this, I feel I will be sticking with the more traditional AIC or BIC techniques of attribute selection.

Estimation of variance of housing prices using spatial conditional heteroskedasticity (SARCH) model with an application to Boston housing price data

This publication from an American team was an economic report regarding the analysis of real estate data and the importance of real estate market analysis. The objective was to model housing price volatility in the context of spatial regression framework. They use a hedonic analysis, which examines the relationship between a commodity and the attributes of the commodity. In the context of real estate, this equates to a regression model in which the value of the home is related to it's attributes. This paper looks at the economic view for housing attributes and their implicit valuation. I like this paper because it gives an economic or market view on how to look at my dataset, which is good to keep in mind when pick attributes.

A novel kNN algorithm with data-driven k parameter computation

This publication from a team of Chinese data analyst investigate the simple, yet popular, kNN classification method. They discuss the methods and techniques in selecting the optimal K value. The group compares the predictive accuracy of kNN to LASSO-kNN and to their proposed S-kNN. Their proposal works to give a more accurate prediction as is replaces a fixed k value with a k value that differs for the test samples according to the distribution.

Exploratory data analysis

This publication looks at exploratory analysis and the various uses it has in data analysis and statistical modeling. This paper references the Book 'Exploratory Data Analysis' by Turkey JW, a book referenced by many young data scientists as a good foundation to learning the tricks and basics of data analysis. This paper talks about various visualization techniques, normality checks, using t-test and correlation test to start the data learning process.

Exploratory data analysis as a foundation of inductive research

This publication from an American team looks at the deductive and inductive approaches of Exploratory data analysis. It discusses logical methods of EDA, which focuses on discovery, exploration and empirical detection. This team also quotes many points from Turkey JW's Exploratory Data analysis book. The paper also shows various univariate and bivariant graphics and how they can be used to gather information about relations. It also offers advice on how to deal with outliers.

The R Project for Statistical Computing

This was not a publication, but this website contains a lot of packages that R includes, as well as good description for different function calls. I did read through man of these to familiarize myself before the project.

Literature Papers Citation

Cock, D. D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education, 19(3). doi:10.1080/10691898.2011.11889627

Wu, Y., Wei, Y. D., & Li, H. (2019). Analyzing Spatial Heterogeneity of Housing Prices Using Large Datasets. Applied Spatial Analysis and Policy. doi:10.1007/s12061-019-09301-x

Wang, X., Jiang, Y., Huang, M., & Zhang, H. (2013). Robust Variable Selection With Exponential Squared Loss. Journal of the American Statistical Association, 108(502), 632-643. doi:10.1080/01621459.2013.766613

Du, J., Chen, X., Kwessi, E., & Sun, Z. (2017). Model averaging based on rank. Journal of Applied Statistics, 45(10), 1900-1919. doi:10.1080/02664763.2017.1401051

Simlai, P. (2014). Estimation of variance of housing prices using spatial conditional heteroskedasticity (SARCH) model with an application to Boston housing price data. The Quarterly Review of Economics and Finance, 54(1), 17-30. doi:10.1016/j.qref.2013.07.001

Cheng, D., Zhang, S., Deng, Z., Zhu, Y., & Zong, M. (2014). KNN Algorithm with Data-Driven k Value. Advanced Data Mining and Applications Lecture Notes in Computer Science, 499-512. doi:10.1007/978-3-319-14717-8_39

Morgenthaler, S. (2009). Exploratory data analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), 33-44. doi:10.1002/wics.2

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. Human Resource Management Review, 27(2), 265-276. doi:10.1016/j.hrmr.2016.08.003

The R Project for Statistical Computing. (n.d.). Retrieved from https://www.r-project.org/

Dataset:

The dataset that I used is the Ames, Iowa dataset. Gathered and provided by Dean De Cock, this dataset serves as an alternative to the Boston Housing Dataset. The author's intention was to create an alternative end year project for his undergrad students. I obtained the data from Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data). The data was presenting in a .csv file. Train and test set was split for us.

There are 46 categorical (23 nominal attributes, 23 ordinal) and 34 quantitative (14 discrete and 20 continuous). Each observation has a unique 'ID' identifier. For our analysis and predictions, the target prediction variable 'SalesPrice'. The data entries to each of the attributes are either char or integer.

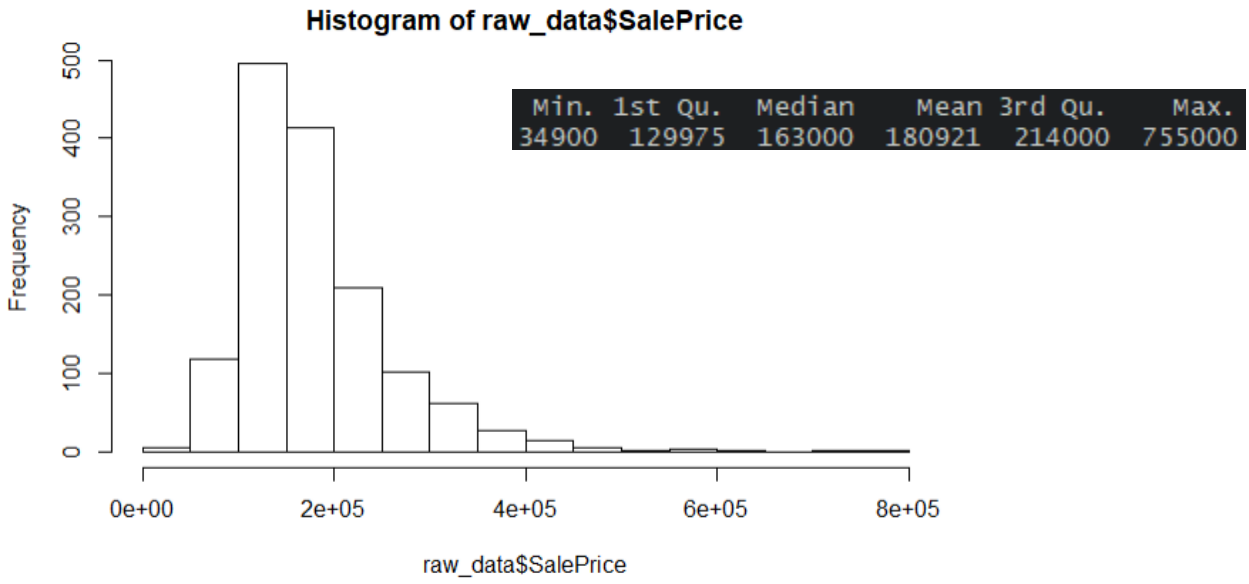
The nominal attributes include: MSSubClass, MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, OverallQual, OverallCond, RoofStyle, RoofMat1, Exterior1st, Exterior2nd, MasVnrType, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Heating, HeatingQC, CentralAir, Electrical, KitchenQual, Functional, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PavedDrive, PoolQC, Fence, MiscFeature, SaleType, SaleCondition.

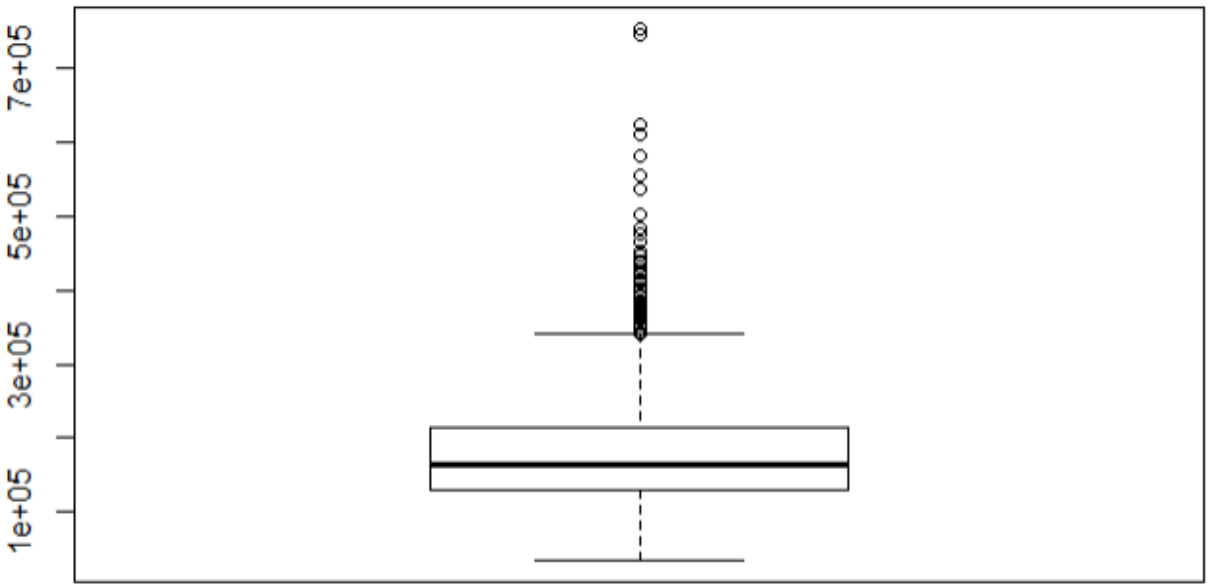
The quantitative variables are: Id, LotFrontage, LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, X1stFlrSF, X2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, X3SsnPorch, ScreenPorch, PoolArea, MiscVal, MoSold, YrSold, SalePrice.

Due to the huge list of variables, I am providing a link to data dictionary which includes all variable information: https://github.com/AbdullahArshad94/HousingData/blob/master/old_data.txt

Initial look:

Looking at our only dependent variable, Sale Prices, we can see some neat initial findings.



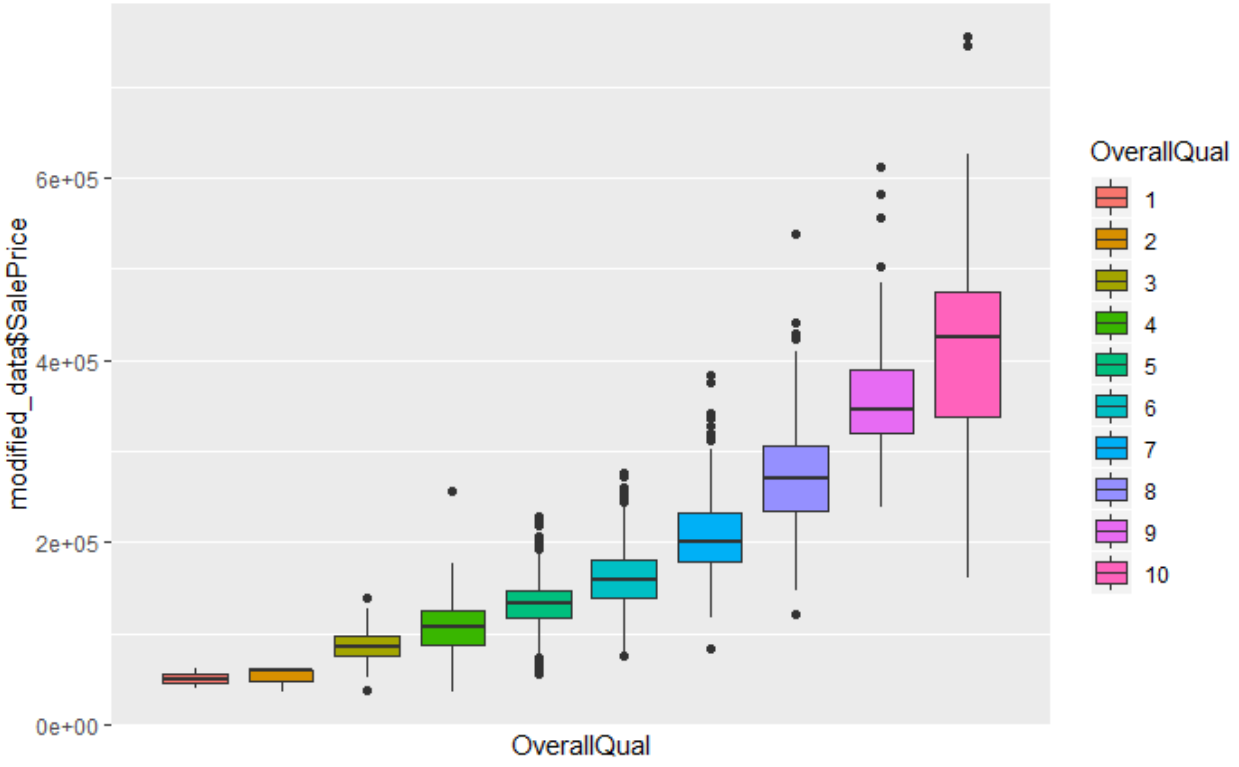


We can see that the average house price in this data set is $180000. We have also figured out that there are 61 outliers in the sale price bracket. As of now, I am not going to do anything against these outliers, perhaps while dealing with outliers in each attribute, or when handling missing values, some of these might disappear.

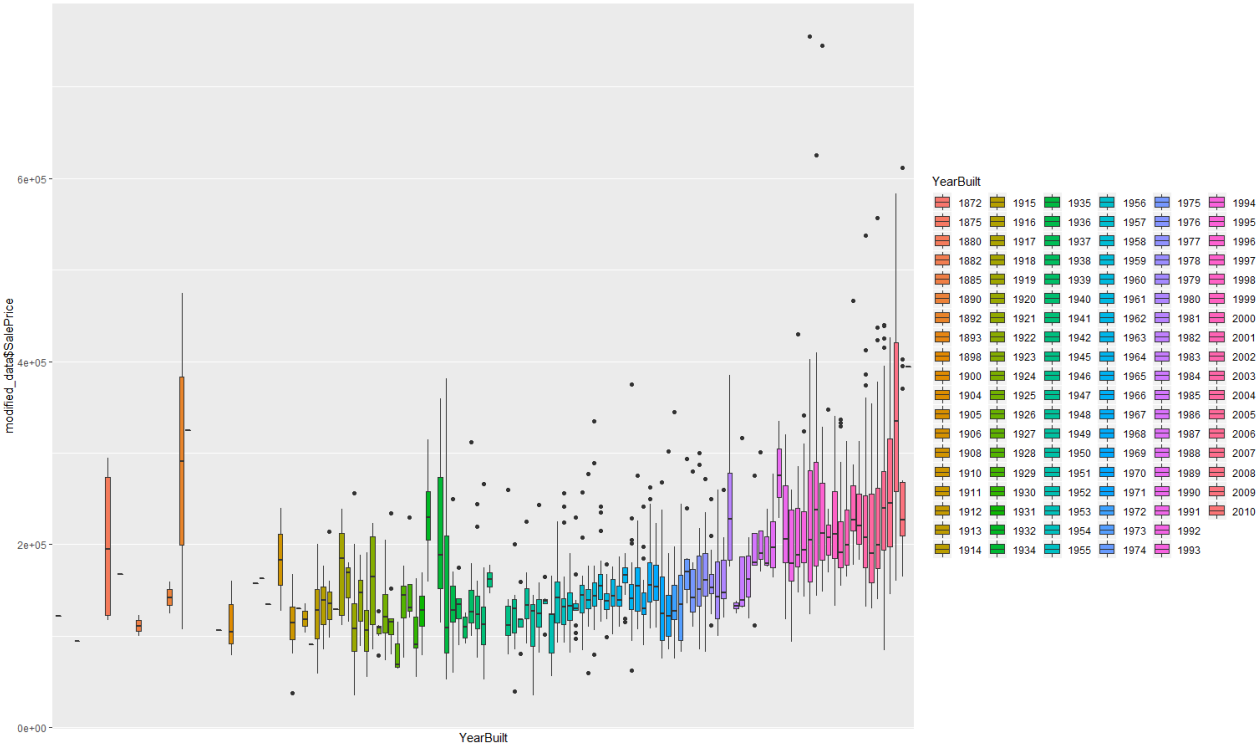
```
---NULL COUNT---
Number of nulls in  LotFrontage : 259
Number of nulls in  Alley : 1369
Number of nulls in  MasVnrType : 8
Number of nulls in  MasVnrArea : 8
Number of nulls in  BsmtQual : 37
Number of nulls in  BsmtCond : 37
Number of nulls in  BsmtExposure : 38
Number of nulls in  BsmtFinType1 : 37
Number of nulls in  BsmtFinType2 : 38
Number of nulls in  Electrical : 1
Number of nulls in  FireplaceQu : 690
Number of nulls in  GarageType : 81
Number of nulls in  GarageYrBlt : 81
Number of nulls in  GarageFinish : 81
Number of nulls in  GarageQual : 81
Number of nulls in  GarageCond : 81
Number of nulls in  PoolQC : 1453
Number of nulls in  Fence : 1179
Number of nulls in  MiscFeature : 1406
```

Initial look also shows us there are a lot of missing values, but a quick assumption will be that many have a meaning to be investigated.

Plotting some of the attributes against the dependent Sales Price gives me some interesting plots.

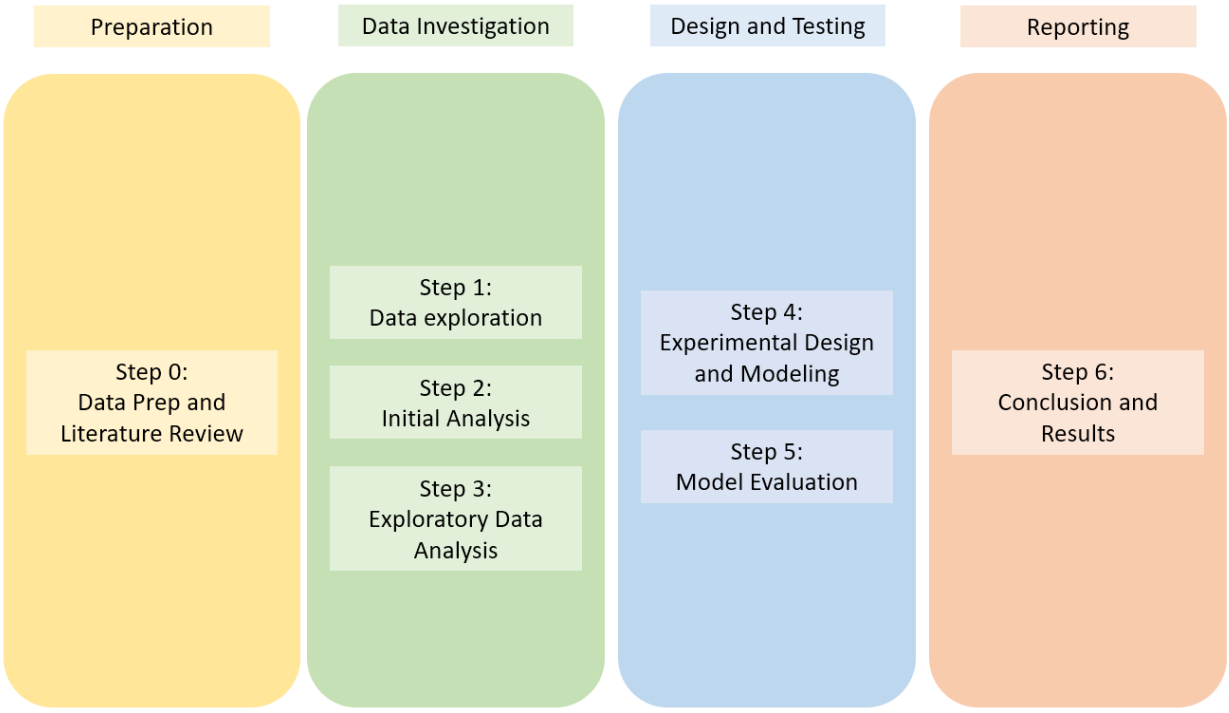


Some of the plots give a general sense of how the dependent variable behaves.



Approach:

My approach to this assignment is divided in four main stages. Each stage may have multiple steps within it, which itself might contain multiple components.



Stage 1: Preparation

In this stage I will start by looking to different journals and articles that will help me in my project. Starting with the journal by the Professor that created the dataset. I will also read articles regarding concepts that I feel that I will need to address in my project, such as; Data Exploration, EDA, Regression, R.

I will also import my data into RStudio's and start the project here. Make sure the dataset is imported properly by running a few checking tests.

Stage 2: Data Investigation

This is the stage where I start to work with my project with R.

In the data exploration step I will be looking through the data and checking the entries for any incorrect entry. Making sure data types are consistent and all attributes are within the data dictionary. I will also be investigating the data dictionary to see whether the entries make sense, and there is no incorrect entry (i.e a numerical value where a categorical entry should be).

In the Initial analysis, I will look to complete my univariant, bivariant, and multivariant analysis. I will start my univariant analysis by cleaning the columns, checking for missing values, looking for and addressing outliers, fixing categorical variables, addressing the distribution of categorical factors. When dealing with missing values, I will be careful to see whether removal is needed or if an alternative can be

assumed. In terms of outliers, I will be looking to see the affect that removing it would have and will assess if the is the correct step. The second part of the initial analysis will be bivariant and multivariant. With the help of correlation tests, I will create pairwise visualization and relations. Correlation analysis will be done here.

For the Exploratory data analysis, I will be working to normalize the data, subset groups, creating decision rules and perform clustering. This is also the step where I perform dimensionality reduction. Because I have 79 exploratory variables, dimensionality reduction will be important and must be comprehensive. I am looking to use low variance filters, PCA and information gain as some tools to make my decision. Regression, as such from my literature, might also be tested and used.

Stage 3: Design and Testing

In this stage I will be using my refined data and machine learning concepts to design a prototype for my design problem.

In the Experimental design and modeling step I will be splitting the data and balancing. I will also be performing multiple modeling techniques, such as; Classification and regression methods.

After constructing the models, I will be moving onto the Evaluation stage. Here I will evaluate each model and pick best performing one. I will use regression and classification evaluation techniques to judge different models. In this stage I will also try to improve the model by addressing overfitting concerns.

Stage 4: Reporting

This final stage will include all documentations; Conclusions and Results. I will be creating a presentation, organizing my weekly milestone reports and creating my final project conclusion report. In this step I will also finalizing all edits and commits onto my GitHub page. This stage is the most important stage, because it will be where I will be addressing the project problem and sharing all that was learnt from the assignment.

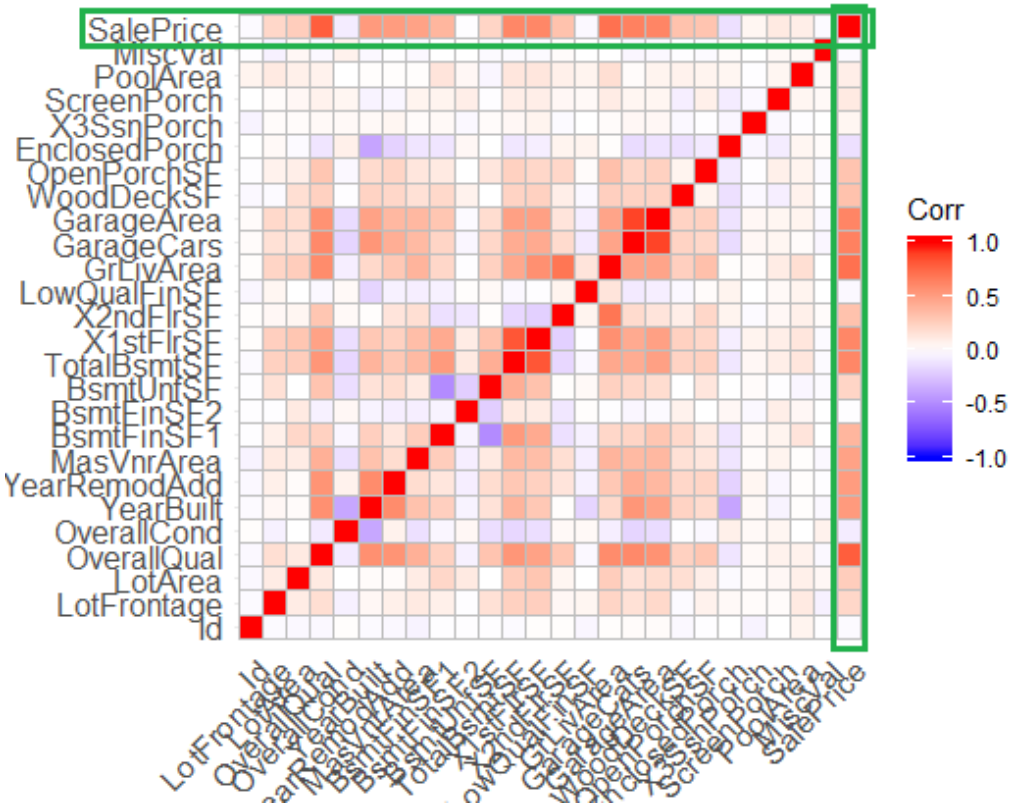
Results

First, I would like to show the effects of our cleaning.

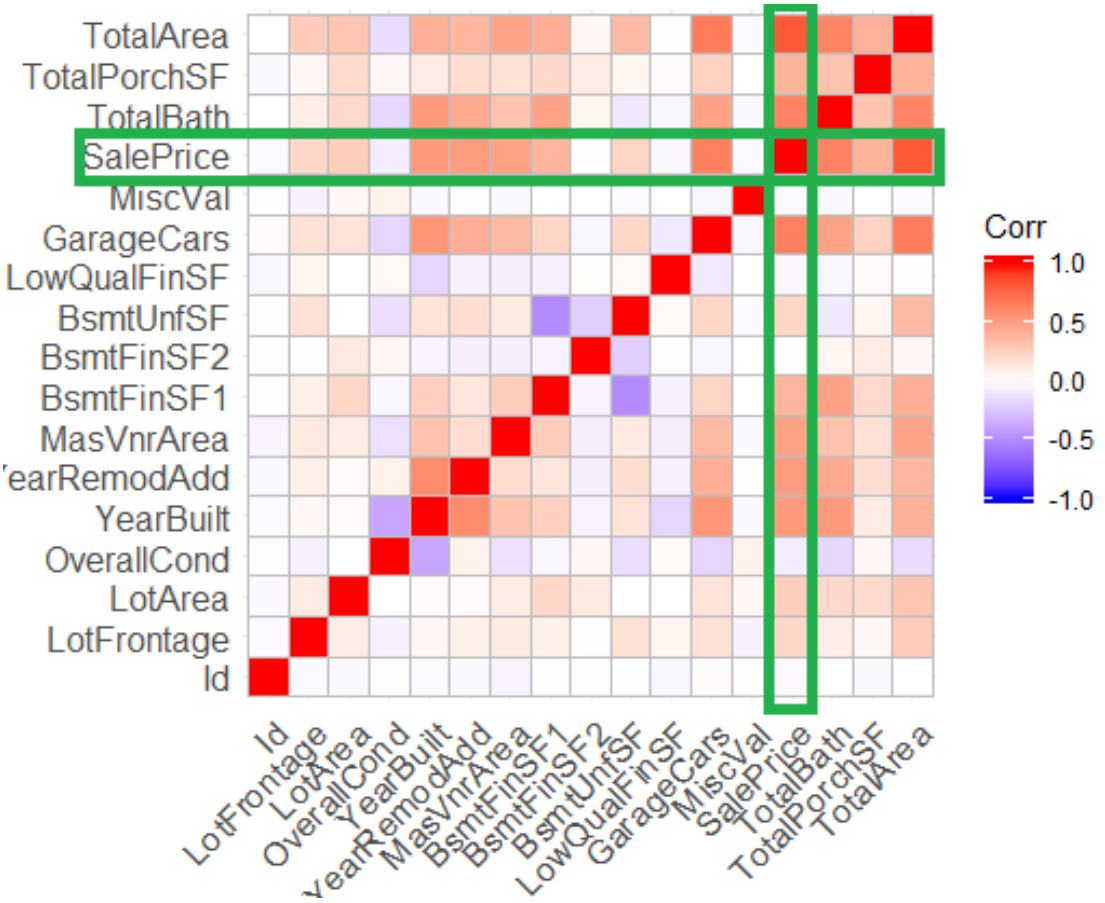
We went from a data set of 1460 observations and 81 attributes to a dataset of 1281 observations and 40 attributes. I only removed about 180 observations, but also got rid of over half our attributes. I gave a high importance to maintaining the size of the dataset, in terms of observations. I wouldn't remove entries unless I deem them necessary. A lot of cleaning and feature selection took place. A lot of the categorical attributes were removed based off the distribution of the variables, in which I created some rule. If there was a categorical attribute in which one level held over 50% of the observations, I would drop it. Or if it had two variables that had over 65% of the observations, I would also drop it. There was no need for such attributes with so little variance in my models. Outliers were also removed if deemed unnecessary, but this was done after all other data cleaning methods. Correlation was used to remove overly correlated variables.

I also did some feature engineering where I decided to combine some attributes into one attribute. For instance, instead of having multiple columns for full and half bathrooms, we decided to add them together. So, a home with one full and one half bathroom would be given the value 1.5. The loss in this is that a home with 1 full bathroom will hold the same weight as a home with 2 half bathrooms, but after consultation, I decided this was an alright addition to my project. Also, instead of having multiple variables hold different areas around the home, I created one for the whole home's area, outdoor area, one for the basement, and one for the land above ground. This helped remove a lot of attributes.

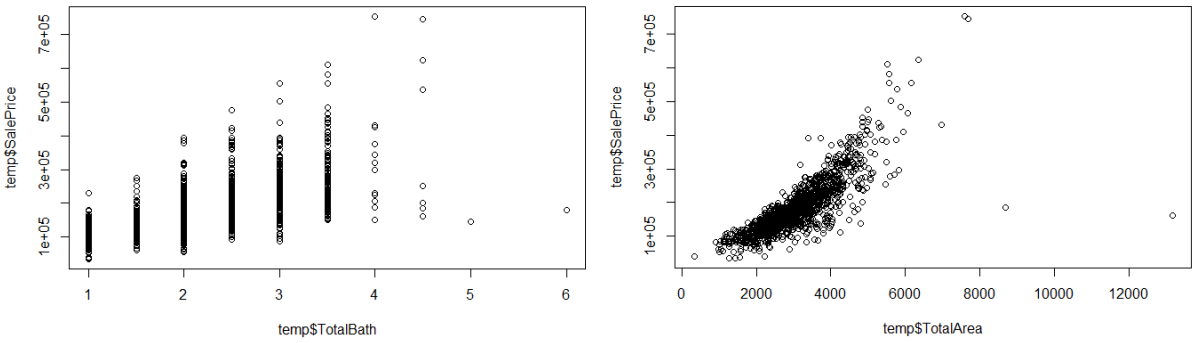
As mentioned, correlation was used to figure out which variables needed to be removed. I have highlighted the independent variable in Green.

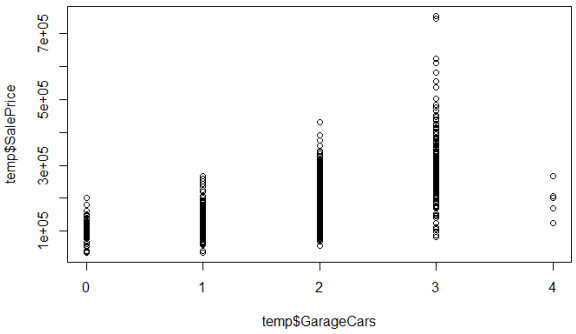


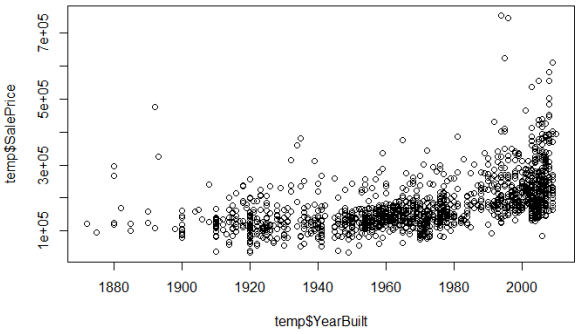
As we can see from the above graphic, there is a very high number of attributes that have over 0.65 correlation to the dependent variable (which I deem too highly correlated). As such, they were either removed or added to create three new variables. The next graphic will show the updated correlation matrix after changes.



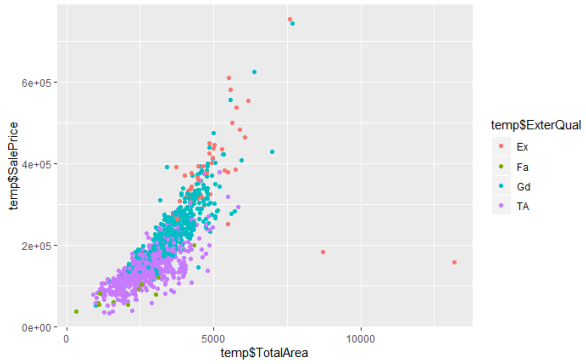
It is important to note that correlation was only done for the numerical data values in the dataset. For the categorical data, I employed other means of measuring relationships. Bi-variant illustrated some very interesting plots that showed correlation between our dependent variable and important independent variables, particularly the categorical variables. More bathrooms or larger homes sell for more.

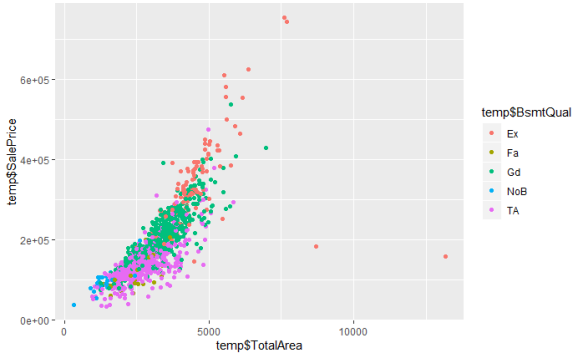


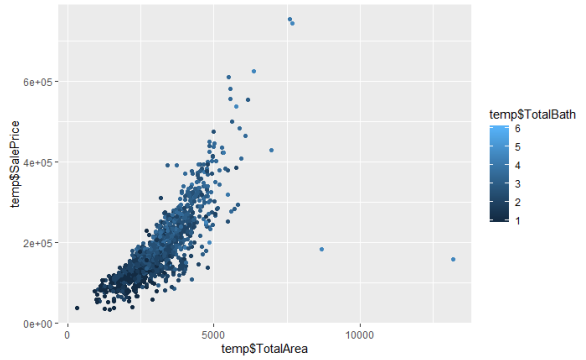


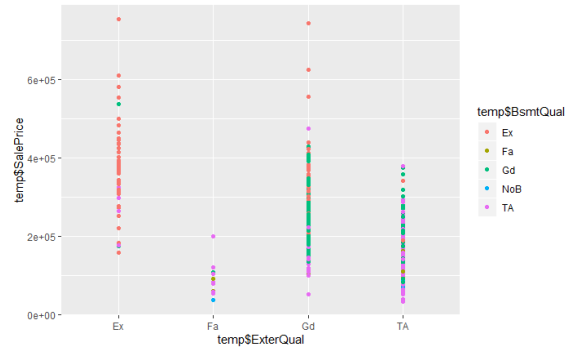


These are some of the more obvious cases of positivly correlated attributes, where we can see that the Sale price would increase as the attribute was gaining favorable values or levels. Newer homes sell for more.





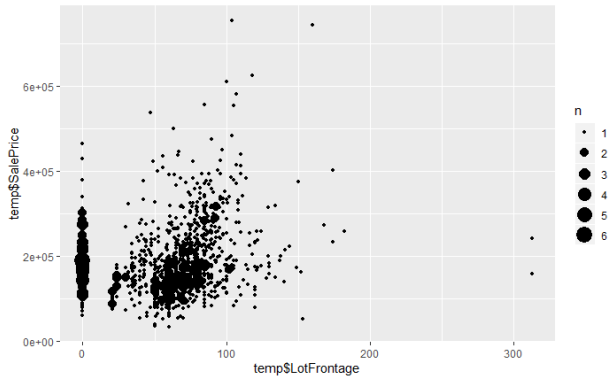




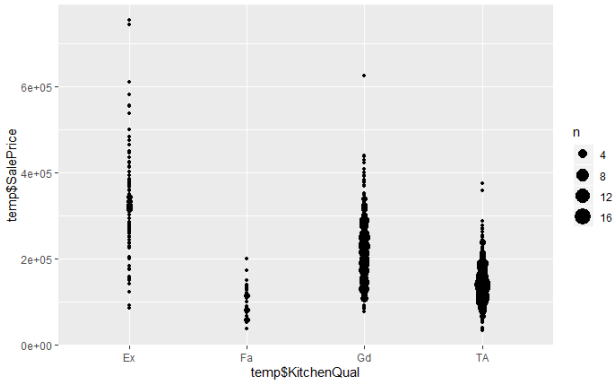
This is even more apparent with multi variant plots. These stories helped illustrate a picture as to why house prices go up, and what factor are more important. However, this information is similar to correlation. But unlike the limitation correlation matrix has, these plots help determing correlation between categorical factors.

Using the Geom count function from the ggplot library, I was able to construct some a counting plot that indicated counts in a pointed area of a graph.

For instance, the following plot helps to show that many of the observations had 0 lotfrontage and still the price was dynamic.



These plots are a bit more helpful because they offer more than a quick plot, as they show a density.



After using all the mention methods, I confident with my data preparation and Exploratory data analysis.

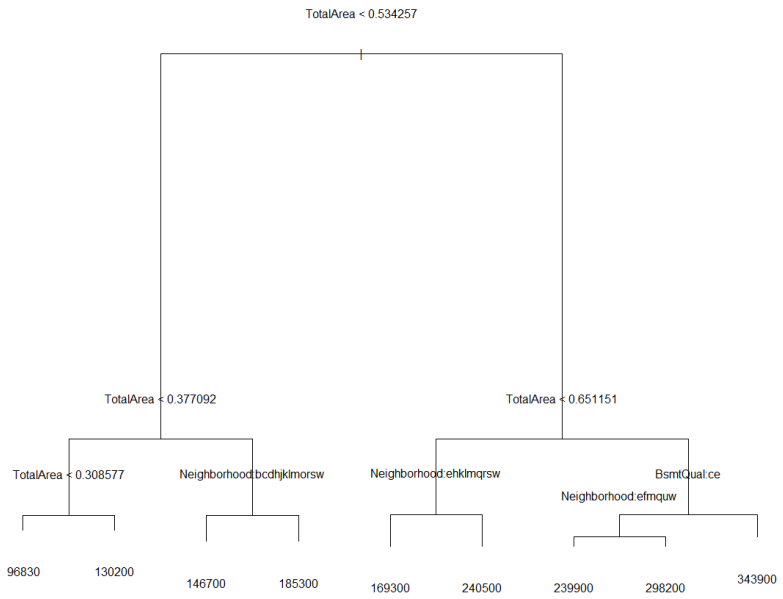
Our final variables were;
MSSubClass,LotFrontage,LotArea,LotShape,Neighborhood,HouseStyle,OverallCond,YearBuilt,YearRemodAdd,RoofStyle,,Exterior1st,Exterior2nd,MasVnrType,ExterQual,,ExterCond,,Foundation,BsmtQual,,BsmtExposure,BsmtFinType1,BsmtFinSF1,BsmtFinType2,BsmtUnfSF,HeatingQC,BedroomAbvGr,KitchenAbvGr,KitchenQual,TotRmsAbvGrd,Fireplaces,FireplaceQu,GarageType,GarageYrBlt,GarageFinish,GarageCars,MoSold,YrSold,TotalBath,TotalPorchSF,TotalArea,SalePrice.

After finishing our Preparation work, we can move forward to Modeling. But before that, I decided to normalize all my numerical data, including the date of construction. I feel that this still hold useful data. I split my data into a training and test set. Training would be 80%, and training would be 20%. I implemented stratified selection, as I ensured at least one factor for each categrorical variable made it to the training data.

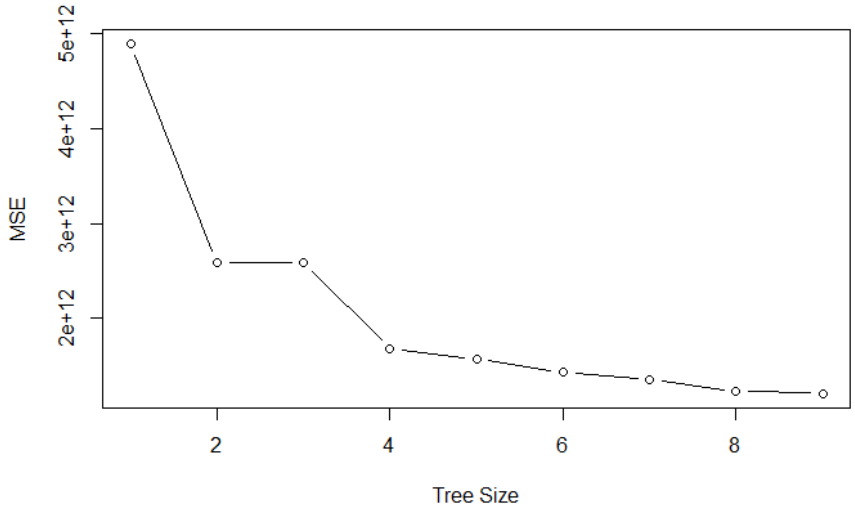
Our first model was Decision Tree, or to be more accurate Regression Tree.

```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 1070 4.885e+12 173600
   2) TotalArea < 0.534257 775 1.156e+12 144600
     4) TotalArea < 0.377092 342 2.677e+11 116900
       8) TotalArea < 0.308577 136 7.569e+10  96830 *
       9) TotalArea > 0.308577 206 1.009e+11 130200 *
     5) TotalArea > 0.377092 433 4.206e+11 166400
      10) Neighborhood: Blueste,BrDale,BrkSide,Edwards,IDOTRR,MeadowV,Mitchel,NAmes,NPkVill,OldTown,Sawyer,SWISU 212 1.088e+11 146700 *
      11) Neighborhood: Blmngtn,ClearCr,CollgCr,Crawfor,Gilbert,NoRidge,NridgHt,NWAmes,SawyerW,Somerst,StoneBr,Timber,Veenker 221 1.510e+11 185300 *
   3) TotalArea > 0.534257 295 1.354e+12 250000
     6) TotalArea < 0.651151 184 4.701e+11 219200
      12) Neighborhood: ClearCr,Edwards,MeadowV,Mitchel,NAmes,NWAmes,OldTown,Sawyer,SWISU 55 7.959e+10 169300 *
      13) Neighborhood: Blmngtn,BrkSide,CollgCr,Crawfor,Gilbert,NoRidge,NridgHt,SawyerW,Somerst,StoneBr,Timber,Veenker 129 1.954e+11 240500 *
     7) TotalArea > 0.651151 111 4.212e+11 301000
      14) BsmtQual: Gd,TA 66 1.690e+11 271700
        28) Neighborhood: ClearCr,CollgCr,NAmes,NWAmes,Somerst,SWISU 30 5.842e+10 239900 *
        29) Neighborhood: Crawfor,Mitchel,NoRidge,NridgHt,OldTown,SawyerW,StoneBr,Timber 36 5.500e+10 298200 *
      15) BsmtQual: Ex 45 1.129e+11 343900 *
```



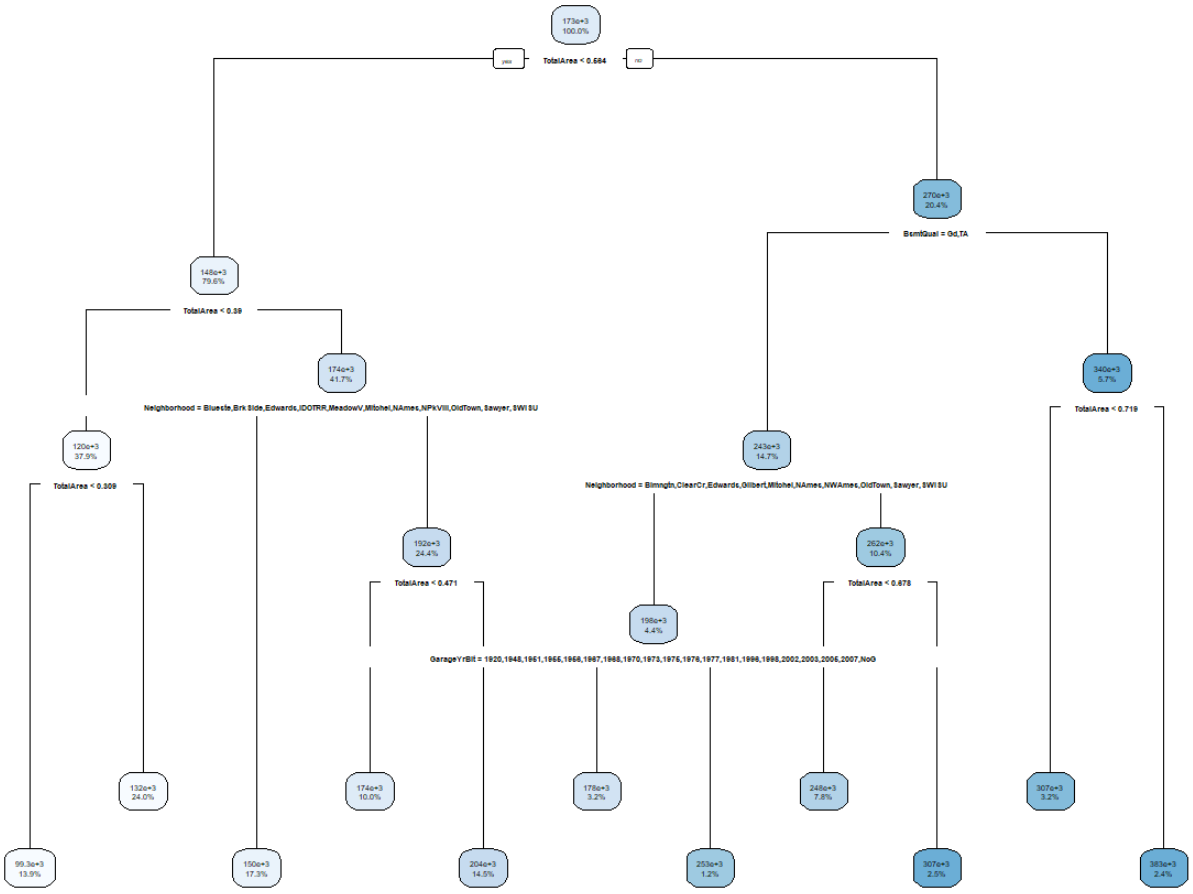
And after pruning, we can see that 9 is a good tree size. We can see that in an error per tree size graph.



However, I wanted to try Decision tree using the rpart library. This one yielded a slightly different tree.

```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 1024 4.860792e+12 173247.40
   2) TotalArea< 0.5644177 815 1.393362e+12 148453.40
     4) TotalArea< 0.3903417 388 2.881968e+11 119810.40
       8) TotalArea< 0.3085774 142 7.187546e+10  99308.34 *
       9) TotalArea>=0.3085774 246 1.221802e+11 131644.90 *
     5) TotalArea>=0.3903417 427 4.975905e+11 174480.30
      10) Neighborhood=Blueste,BrkSide,Edwards,IDOTRR,MeadowV,Mitchel,NAmes,NPkVill,OldTown,Sawyer,SWISU 177 8.668870e+10 149951.30 *
      11) Neighborhood=Blmngtn,ClearCr,CollgCr,Crawfor,Gilbert,NoRidge,NridgHt,NWAmes,SawyerW,Somerst,StoneBr,Timber,Veenker 250 2.290061e+11 191846.90
        22) TotalArea< 0.4706241 102 3.551035e+10 173879.60 *
        23) TotalArea>=0.4706241 148 1.378742e+11 204229.80 *
   3) TotalArea>=0.5644177 209 1.012702e+12 269932.00
     6) BsmtQual=Gd,TA 151 4.180082e+11 243116.00
      12) Neighborhood=Blmngtn,ClearCr,Edwards,Gilbert,Mitchel,NAmes,NWAmes,OldTown,Sawyer,SWISU 45 1.316607e+11 198185.80
        24) GarageYrBlt=1920,1948,1951,1955,1956,1967,1968,1970,1973,1975,1976,1977,1981,1996,1998,2002,2003,2005,2007,NoG 33 6.463080e+10 178342.40 *
        25) GarageYrBlt=1950,1961,1972,1974,1988,1997,1999,2006 12 1.830219e+10 252755.10 *
      13) Neighborhood=CollgCr,Crawfor,NoRidge,NridgHt,SawyerW,Somerst,StoneBr,Timber,Veenker 106 1.569400e+11 262190.10
        26) TotalArea< 0.6778243 80 6.175055e+10 247555.30 *
        27) TotalArea>=0.6778243 26 2.533450e+10 307220.30 *
     7) BsmtQual=Ex 58 2.034161e+11 339746.20
      14) TotalArea< 0.7185321 33 6.234175e+10 306812.20 *
      15) TotalArea>=0.7185321 25 5.803337e+10 383219.20 *
```



This tree is much larger, as the size is now 11. The illustration is much better. We can clearly see by the color that as you move toward the right of the chart, the Sale Price increase.

Looking back at both the trees, it is evident that Total Area is the most important feature in determining price. Perhaps I should have removed it because of it's high correlation, but I decided to leave it in due to its higher information gain and also because it is an addition of many other variables.

The next model I investigated was Random forest.

```
 randomForest(formula = rdata_train$SalePrice ~ ., data = rdata_train,      proximity = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 12

          Mean of squared residuals: 528675042
                    % Var explained: 88.42
```

Check the error rate, increasing the number of trees did not significantly change the mean of squared residuals.

Finally, the last model was the Elastic Net Regression algorithm. The interesting part of this type of regression is that it incorporated Ridge and Lasso regression. Both perform slightly different depending on the data conditions, so it gives the user room to pick which method they'd like to use more of. Elastic net regression has an alpha value that helps decide which part of the spectrum the user wishes to use,

Ridge or Lasso. I built a loop that worked to look for the best alpha value for my data's prediction RMSE. After testing, alpha = 0.6 (slightly more LASSO than RIDGE) gave us the lowest error rate.

Alright, so we've set up all our models. Now it is time to compare them and decide which one performs the best. Before that we must figure out our metric in deciding an algorithm's performance. For regression, we would use error calculations. R-Squared, Standard Error, AIC/BIC, RMSE, MAE are all ways to rank algorithms. I've decided to use RMSE for my models.

```
[1] "For Regression Models RMSE and MAE are commonly used when comparing regression models"
                  Model     RMSE      MAE
1        Decision Tree 1 29739.19 22090.49
2        Decision Tree 2 32019.63 23595.01
3          Random Forest 19590.35 13766.77
4 Elastic Net Regression 23397.19 17545.21
```

As we can see, Random Forest performs the best. It is also worth mentioning that the Random forest was run on 10-fold cross-validation. We can also see that even though the second Decision tree method was better illustrated and had larger size, the error rate is higher. I will be using the first model, going forward.

Finally, the initial outlier from the Sale Prices were removed. I initially wanted to mask them until the end to see how much of an effect they had on the prediction system. Knowing this for the previous graphic, we should also look at MAE as it is more robust than MSE.

```
                  Model      RMSE       MAE
1          Decision Tree  8956.769  7544.963
2          Random Forest  6355.614  3662.325
3 Elastic Net Regression 20157.075 14548.459
```

We can see a dramatic decrease in the error parameter. We can safely say that Random Forest performs the best and results in the lowest error rates. Now that outliers have been removed, we can use RMSE as a means of measurement. It yields an error of approximately 3.5% if held against the average house price.

Conclusion

Through my project I learned to prepare my data by cleaning and addressing concerns in it. Exploring the many stories it holds, and determine what stories and information is necessary for the task of predicting Sale Prices of a home. After a pain staking journey of cleaning the raw data, I had cut my attributes in half and eliminated almost 15% of my observations. Attributes were removed based on high correlation, low variance, % of outlier and if another attribute brought similar information. I divided my data into a test and training sets, while doing so, I implemented stratified selection so that at least each factor was mentioned in the training set. I then ran it through various algorithms. After going through Regression tree, Regression Forest and Net Elastic Regression; I found that that regression tree performed the best. The metric I used to evaluate each algorithm was Root mean squared error, which is preferred for regression problems. The error rate was approximately 3.5% the mean Sales Price.

My code and documentation can be obtained from my GitHub page:
https://github.com/AbdullahArshad94/HousingData