STAT 847: Final Project

DUE: Friday April 19, 2024 by 11:59pm Eastern

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark.

Total word count target is 1000 words, there are 100 marks. Approximately 15 for each of 6 parts (if one part is extra large, it will count for more points), and 10 points for presentation as a whole. Use your visualization and writing skills, as well as your data science skills.

Find a dataset we didn't cover in class, it has to be large enough with enough features to complete the parts listed below. (Notice that 4 are mandatory, 4 are optional)

1) **MUST BE INCLUDED** Describe and justify two different topics or approaches you might want to consider for this dataset and task. You don't have to use these tasks in the actual analysis.

For this project, I will be analyzing the House Prices dataset from the "KAGGLE GETTING STARTED PREDICTION COMPETITION," which includes 1460 observations (houses) and 81 variables suitable for explanatory analysis or feature engineering. I plan to explore two key topics with this dataset:

1. **Identification of Key Features Influencing Home Prices**: I aim to pinpoint the variables that most significantly impact the pricing of a home. Understanding these key drivers will not only aid home developers in tailoring their constructions to meet market demands but will also enable them to command higher prices by emphasizing the features that prospective buyers value most.
2. **Home Price Prediction Model**: By using the identified important features, I will develop a predictive model for home prices. The model will be evaluated to ensure accuracy. This approach is particularly useful for various stakeholders. For realtors and investors, accurately predicting home prices helps in identifying whether homes are underpriced (indicating potential bargains) or overpriced (which could lead to prolonged market listings and decreased attractiveness).

These approaches are not just academic exercises; they have practical implications that could lead to more informed decision-making in the real estate market. By leveraging data to identify and predict key factors affecting home prices, stakeholders can optimize their strategies to maximize returns and efficiency in the property market.

2) **MUST BE INCLUDED** Give a ggpairs plot of what you think are the six most important variables. At least one must be categorical, and one continuous. Explain your choice of variables and the trends between them.

```r
library(GGally)

Q2_plot <- ggpairs(Q2_df)

Q2_plot
```

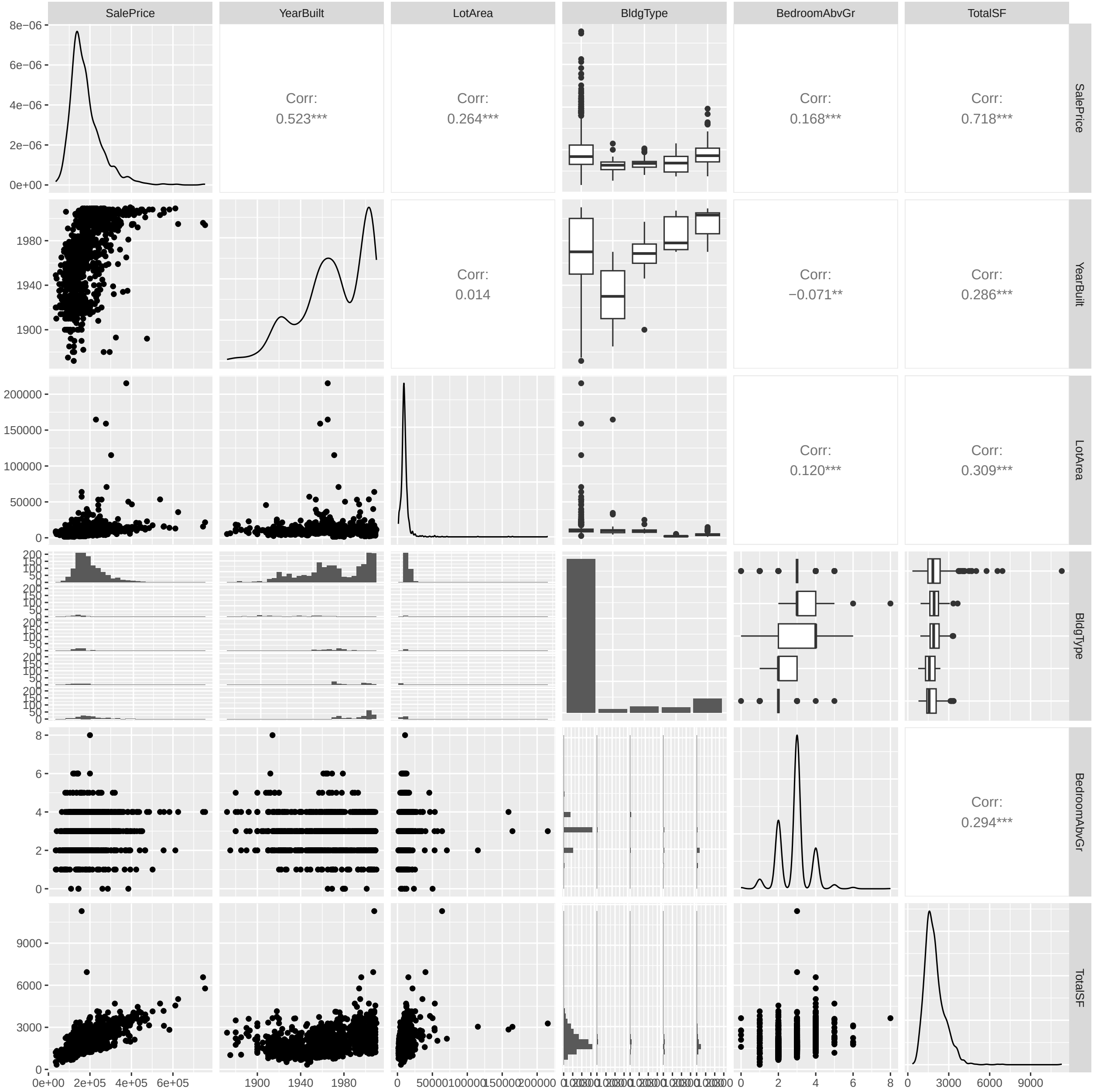


Figure 1: GGpairs plot of the 6 variables I believe are most important.

For this exercise I chose the variables SalePrice, YearBuilt, LotFrontage, LotArea, BldgType, BedroomAbvGr.

1. **SalePrice**: This variable captures the price the home was sold for. This will be my dependant in my prediction model.
2. **YearBuilt**: The YearBuilt variable depicts the original construction date. This is important because housing has definately improved over time and so the construction date can be a predictor for the quality of the build.
3. **TotalSF**: I created this variable to capture the total square footage of the home by adding the square footage of the basement and the square footage of all home area above grade (above soil level). I believe this variable is important because typically bigger families need more home area and are willing to pay more for that so this can definately impact the house price.
4. **LotArea**: This varable captures the lot size in square feet. This is an important factor since the amount of area of a property could increase the value as the home square feet may not capture the amount of land in the transaction.
5. **BldgType**: The variable BldgType captures the type of dwelling:

- 1Fam: Single-family Detached

- 2FmCon Two-family Conversion; originally built as one-family dwelling
- Duplx: Duplex
- TwnhsE: Townhouse End Unit
- TwnhsI: Townhouse Inside Unit

I chose this variable due to the type of house having a great impact on the price. An example of this is that a duplex and a single family detached could have the same square footage however since the duplex can accomodate two families it could be worth more.

6. **BedroomAbvGr**: The BedroomAbvGr variable captures the number of bedrooms above grade (does NOT include basement bedrooms). This variable is important because the number of bedrooms can be a deciding factor for a family on whether they will buy the house or not depending on whether it has enough rooms to accommodate their family.

With regards to the relationships seen in between the variables in the ggpairs plot. The first one that stuck out to me was the strong positive correlation between Total square footage and Sale Price which makes sense since individuals will pay more for a home with more square footage.We observe a weak positive correlation between the number of bedrooms above ground and the total square footage, this is certainly due to the opportunity to build more bedrooms as the square footage increases however its weak since theres only so many bedrooms that a house can have before its too much. we see a weak positive correlation between YearBuilt and total square footage which weeakly tells us that over time homes have been increasing in size. We see little to no correlation to lot area and year built with along with the last correlation tells us that over time more home are being built with square footage in mind over the lot area, which is an important factor to look out for in my model. Im expecting now to see a high importance on total square footage of a home which determined the switch to prioritizing total square footage. Another interesting correlation is the correlation between YearBuilt and sale price, we see a exponential rise that over time is reaching a limit, i believe this is due to the inflation of the dollar and the housing market reaching a peak around 2008 before the housing crisis then we see that the price levels off due to the market crash.

3) **MUST BE INCLUDED** Build a classification tree of one of the six variables from the last part as a function of the other five, and any other explanatory variables you think are necessary. Show code, explain reasoning, and show the tree as a simple (ugly) plot. Show the confusion matrix. Give two example predictions and follow them down the tree.

For this question i will be adding the variables Fireplaces, GarageType, PoolArea, SaleType, and remodel to the dataset. Fireplace is a count of the number of fireplaces in the home, I am adding this variable since personally I am a big fan of fireplaces and would need a fireplace in my home for me to make the purchase. GarageType is a variable that captures what type of garage the home has, again I am choosing to add this variable due to me enjoying spending time in my garage working on my car so I would like to see the impact the GarageType could have on a home. PoolArea captures the area of a pool and is 0 if there is no pool, I am adding this to see the impact a pool could have on a home, I am expecting little to no impact. SaleType is a catagorical variable with many factors however the two I am most interested in and most prominent are WD: Warranty Deed - Conventional and New: Home just constructed and sold. and finally remodel is a variable I created that checks if there is a remodel date, if yes then the house was remodeled and it is a 1 else it is 0 so the variable captures if the house was remodeled after being built or not.

I will create a classification tree to predict the number of bedrooms above grade as a function of all the other variables in the dataset (all the variables mentioned in Q2 plus the ones mentioned above).

```r
Q3_df <- house_df %>%
    select(SalePrice, YearBuilt, LotArea, BldgType,
        BedroomAbvGr, GrLivArea, BsmtFinSF1, Fireplaces,
        GarageType, PoolArea, SaleType, YearRemodAdd,
        YearBuilt) %>%
    mutate(TotalSF = BsmtFinSF1 + GrLivArea, remodel = if_else(YearRemodAdd !=
        YearBuilt, 1, 0), GarageType = if_else(is.na(GarageType),
        "NoGarage", GarageType)) %>%
    select(-GrLivArea, -BsmtFinSF1, -YearRemodAdd)

Q3_df$BedroomAbvGr <- as.factor(Q3_df$BedroomAbvGr)
```

```
tree_model <- tree(BedroomAbvGr ~ ., data = Q3_df)

# Plot the tree
plot(tree_model, uniform = TRUE, main = "Classification tree")
text(tree_model, use.n = TRUE, pretty = 0)
```

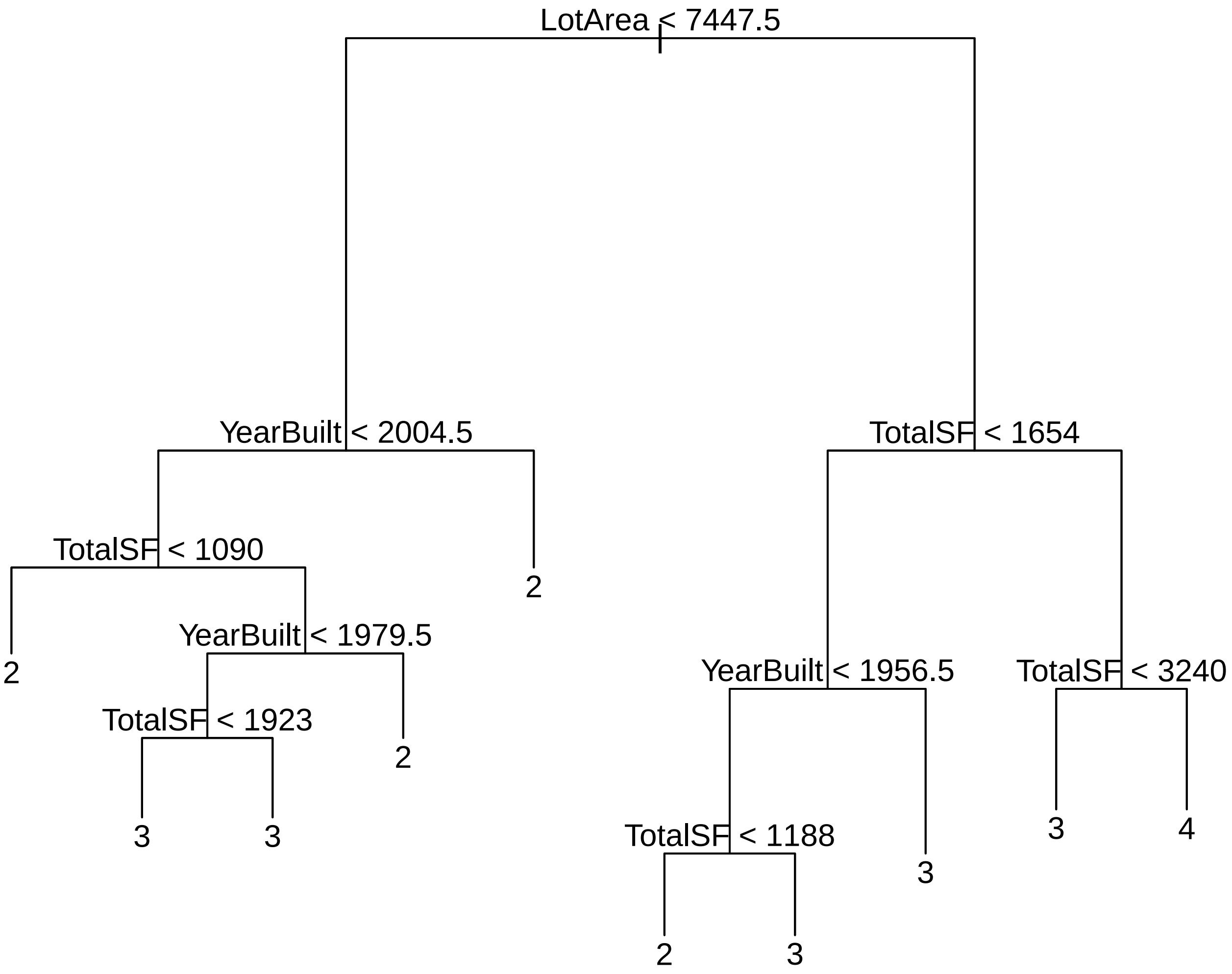


Figure 2: Simple(ugly) plot of the classification tree created to predict number of bedrooms above grade(above soil level).

```r
plt <- as.data.frame(confusion_matrix$table)
plt$Prediction <- factor(plt$Prediction, levels = rev(levels(plt$Prediction)))

ggplot(plt, aes(x = Reference, y = Prediction,
    fill = Freq)) + geom_tile() + geom_text(aes(label = Freq),
    vjust = 1) + scale_fill_gradient(low = "white",
    high = "#009194") + labs(x = "Reference",
    y = "Prediction", title = "Confusion Matrix Visualization") +
    scale_x_discrete(labels = c("0", "1", "2",
        "3", "4", "5", "6", "8")) + scale_y_discrete(labels = c("8",
    "6", "5", "4", "3", "2", "1", "0"))
```

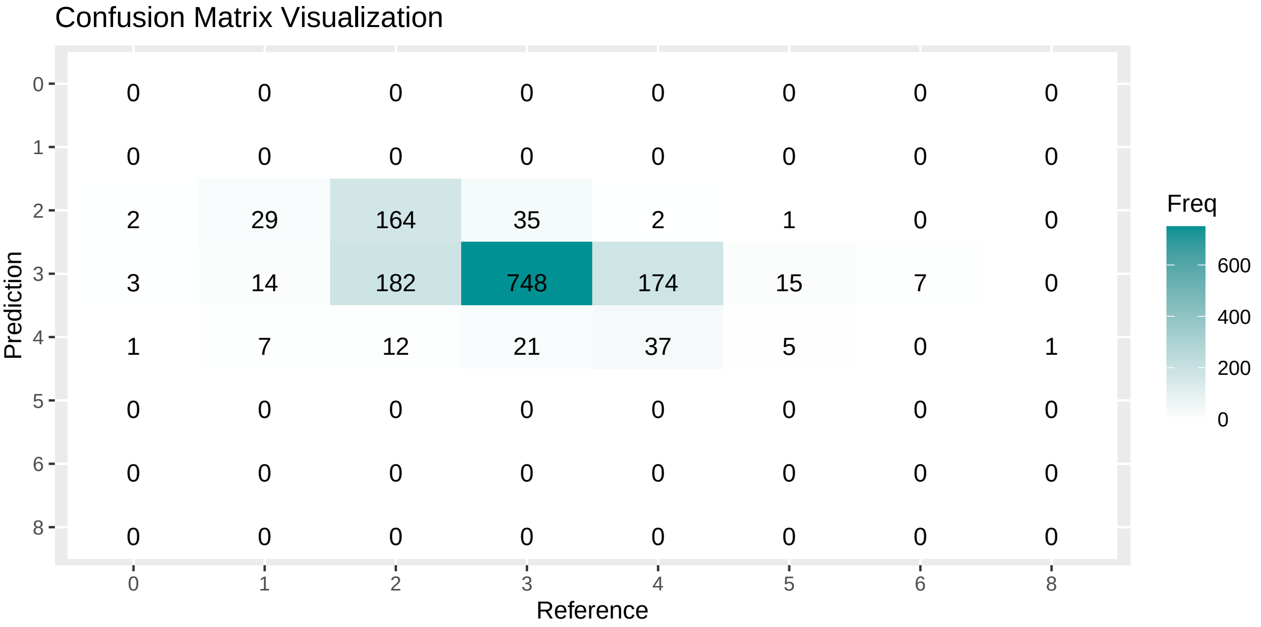


Figure 3: Confusion matrix plot of the predictions made on the housing dataset.

Table 1: Sample Real Estate Data

SalePrice	YearBuilt	LotArea	BldgType	BedroomAbvGr	Fireplaces	GarageType	PoolArea	SaleType	TotalSF	Remodel
90000	1967	10791	Duplex	2	0	CarPort	0	WD	1296	0
118000	1939	7420	2fmCon	2	2	Attchd	0	WD	1928	1

For the first row in the 2 example predictions we first see that the LotArea is 10791 which is greater than 7447.5 therefore since the first decision is false we follow the left tree. Next we see that the YearBuilt is 1967 which is less than 2004.5 therefore that decision tree is true and we follow the right tree leading us to the correct prediction that the number of bedrooms above grade is 2.

For the second row in the 2 example prediction we see that the LotArea is 7420 which is less than 7447.5 since this is true we follow the right tree. Next we see that the TotalSF is 1928 which is greater than 1654 making this decision false so we follow the left tree, next we see tha the YearBuilt less than 1956.5 is true therefore we incorrectly predict that this home has 3 above grade bedrooms when in reality it only has 2.

See the following two pages for visual depictions of the predictions.

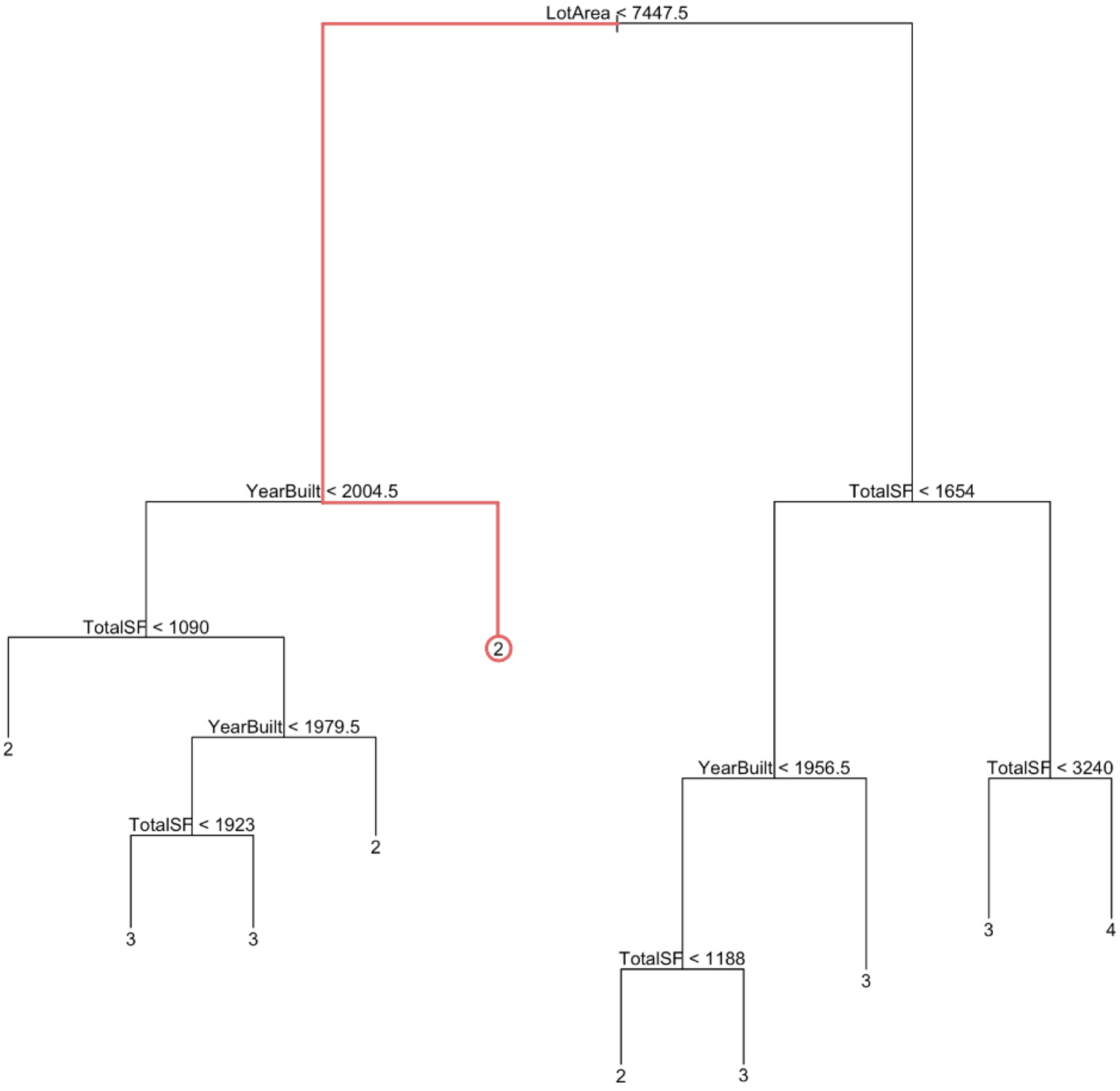


Figure 4: A visual depiction of the tree classification for the first row of the sample real estate data from Table 1.

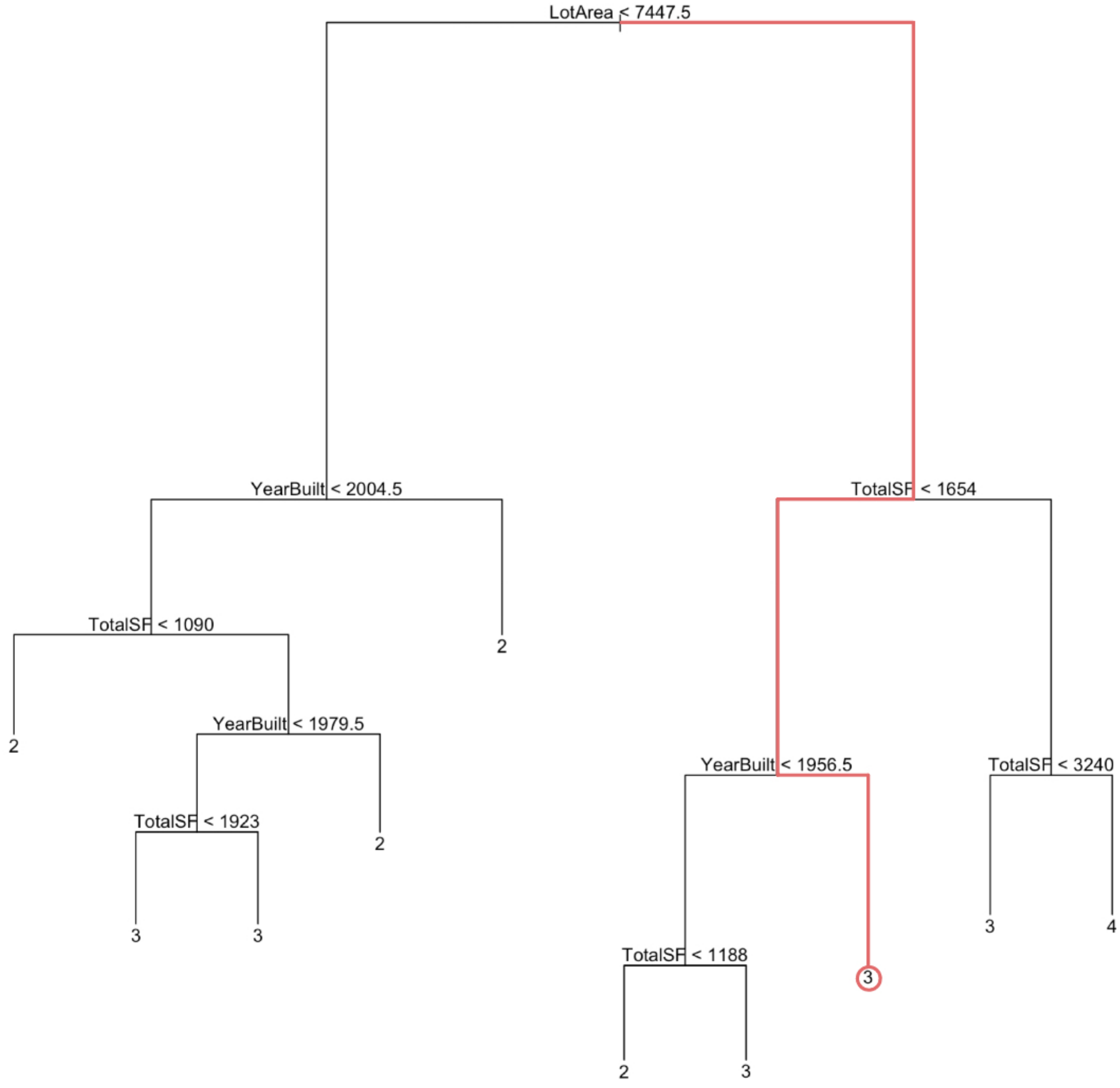
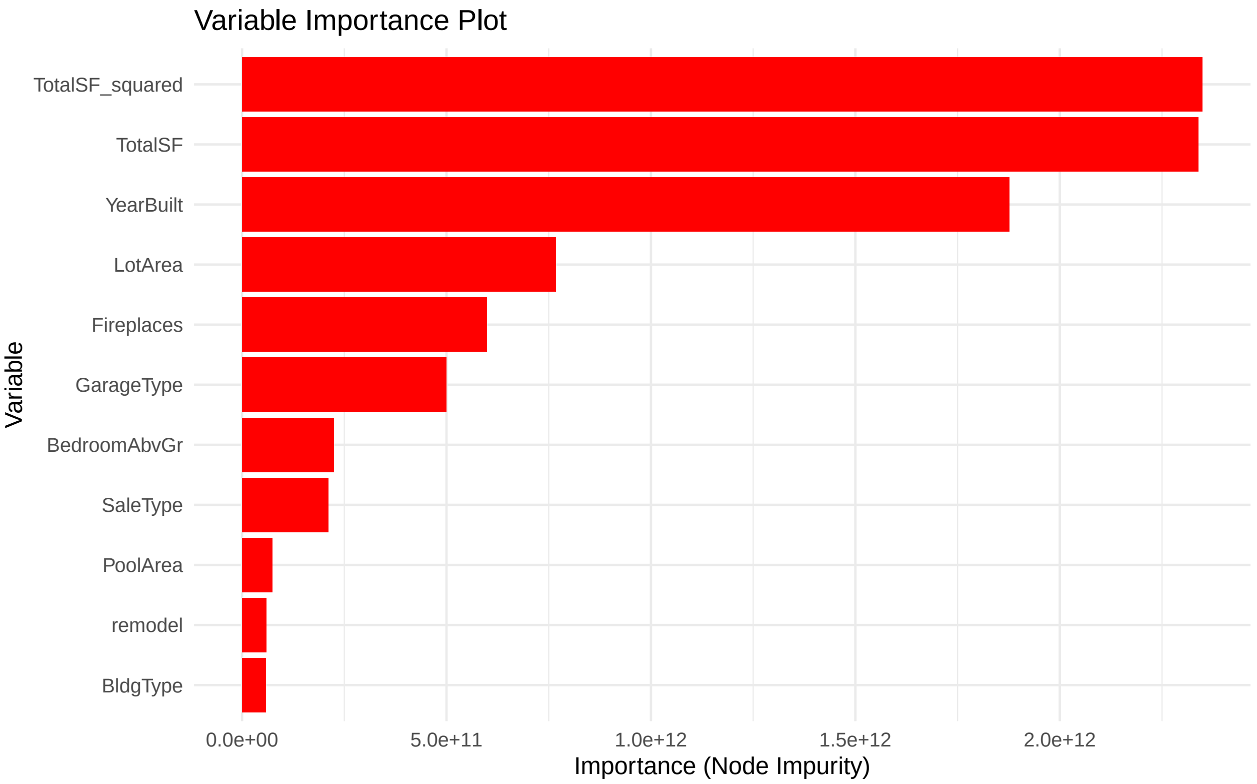


Figure 5: A visual depiction of the tree classification for the second row of the sample real estate data from Table 1.

4) **MUST BE INCLUDED** Build another model using one of the continuous variables from your six most important. This time use your model selection and dimension reduction tools, and include at least one non-linear term.

```r
Q4_df <- Q3_df %>%
    mutate(TotalSF_squared = TotalSF^2)
```

```r
set.seed(21108082)

prelim_rf <- randomForest(SalePrice ~ ., data = Q4_df,
    ntree = 100)

var_importance <- importance(prelim_rf)

importance_df <- as.data.frame(var_importance)
importance_df$Variable <- rownames(importance_df)

importance_df <- importance_df %>%
    select(Variable, IncNodePurity) %>%
    arrange(desc(IncNodePurity))

ggplot(importance_df, aes(x = reorder(Variable,
    IncNodePurity), y = IncNodePurity)) + geom_bar(stat = "identity",
    fill = "red") + coord_flip() + labs(x = "Variable",
    y = "Importance (Node Impurity)", title = "Variable Importance Plot") +
    theme_minimal()
```



```r
set.seed(21108082)
final_rf <- randomForest(SalePrice ~ TotalSF +
    TotalSF_squared + YearBuilt + LotArea + BedroomAbvGr +
    GarageType, data = Q4_df, ntree = 500)

reprtree:::plot.getTree(final_rf)
```

```
## Registered S3 method overwritten by 'reprtree':
##   method    from
##   text.tree tree
```

```
## Loading required package: plotrix
```

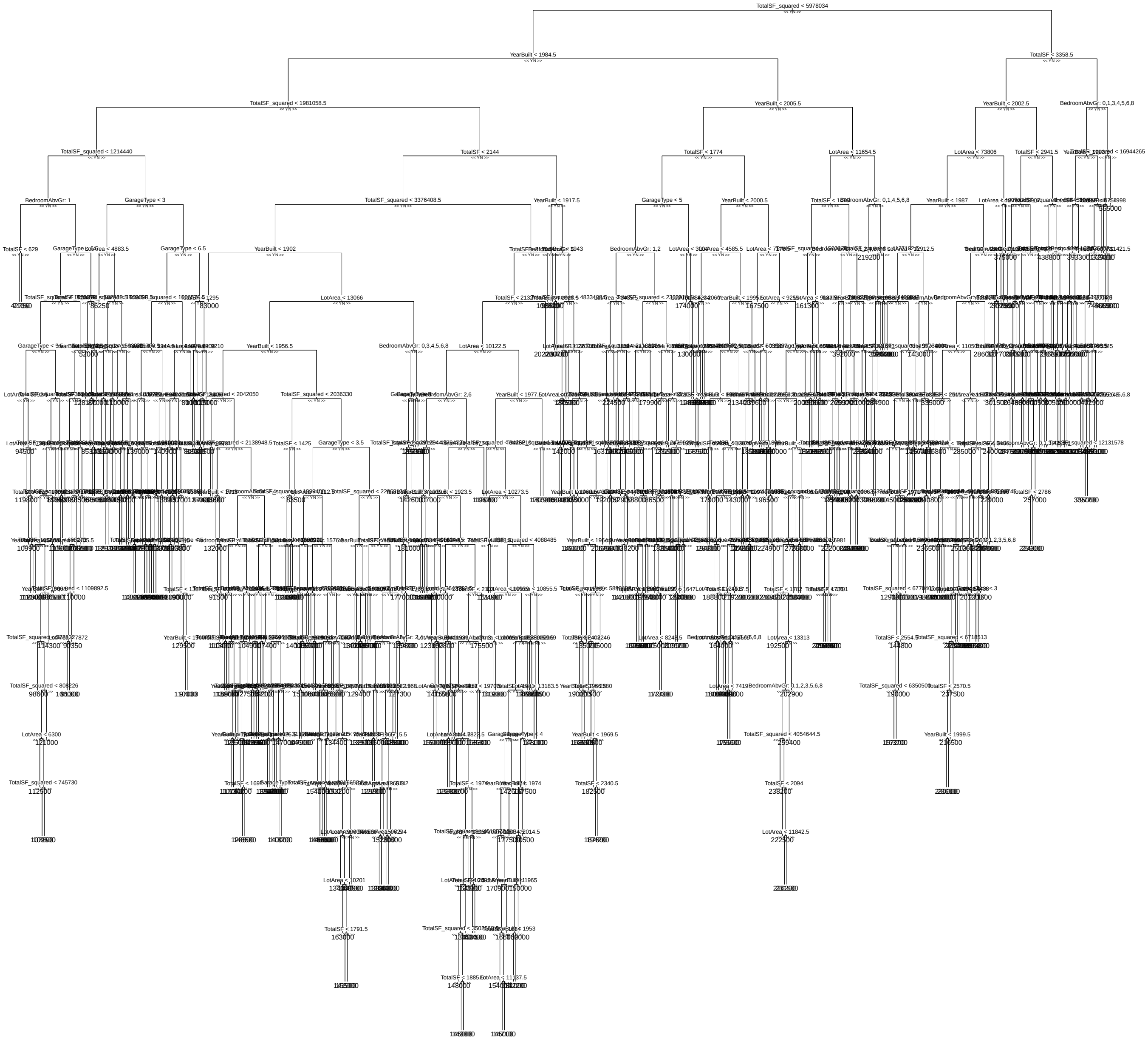


Table 2: Evaluation parameters of our Random Forest model

	Random Forest Model
RMSE	17,356.16
R Squared	0.96

```r
library(ggplot2)
ggplot(Q4_df, aes(x = SalePrice, y = predictions)) +
    geom_point(alpha = 0.5) + geom_smooth(method = "lm",
    color = "blue") + labs(title = "Actual vs. Predicted SalePrice",
    x = "Actual SalePrice", y = "Predicted SalePrice")
```

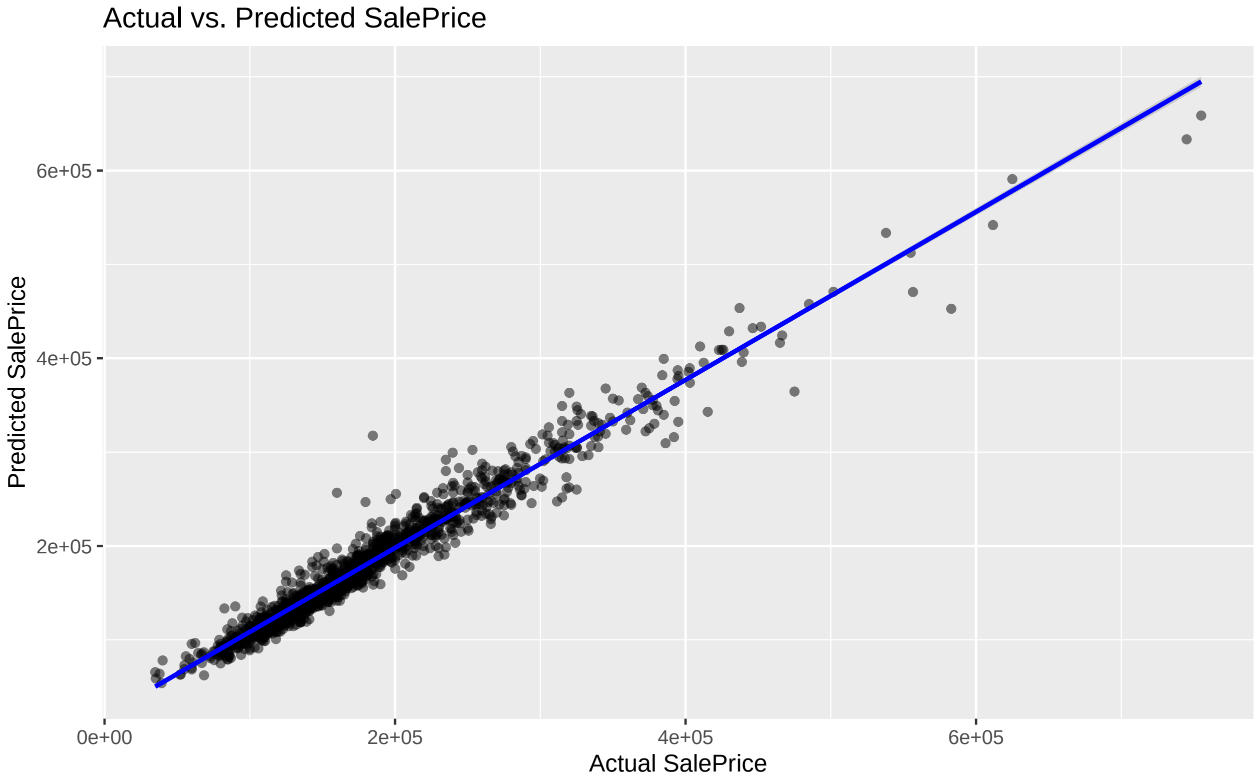


Figure 6: Plot displaying the predicted prices against the actual values with a blue 45 degree line

```r
# Plot t-SNE results
plot(tsne_results$Y[, 1], tsne_results$Y[, 2],
     col = Q4_df$BldgType, pch = 19, cex = 0.5,
     main = "t-SNE plot")
legend("topright", legend = levels(factor(Q4_df$BldgType)),
     col = 1:length(levels(factor(Q4_df$BldgType))),
     pch = 19)
```

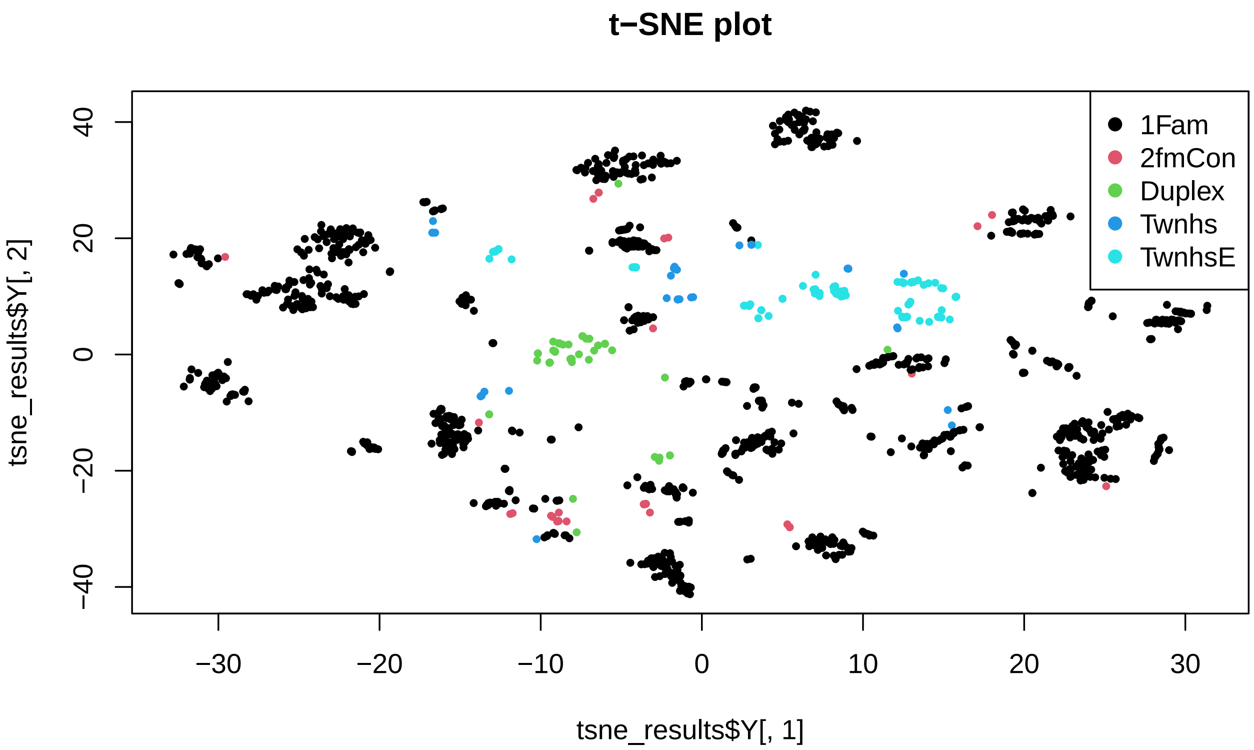


Figure 7: Plot displaying the t-SNE results

6) **OPTIONAL: PICK 2 OF 4** Build a visually impressive ggplot to show the relationship between at least three variables.

```
Q4_df$BedroomAbvGr <- as.factor(Q4_df$BedroomAbvGr)
Q4_df$remodel <- as.factor(Q4_df$remodel)

ggplot(Q4_df, aes(x = BedroomAbvGr, y = SalePrice,
    color = remodel)) + geom_boxplot(outlier.colour = "black",
    outlier.shape = 18, outlier.size = 1, notch = FALSE) +
    labs(title = "A boxplot depicting the relationship between Sale Price, Bedrooms and remodel")
```

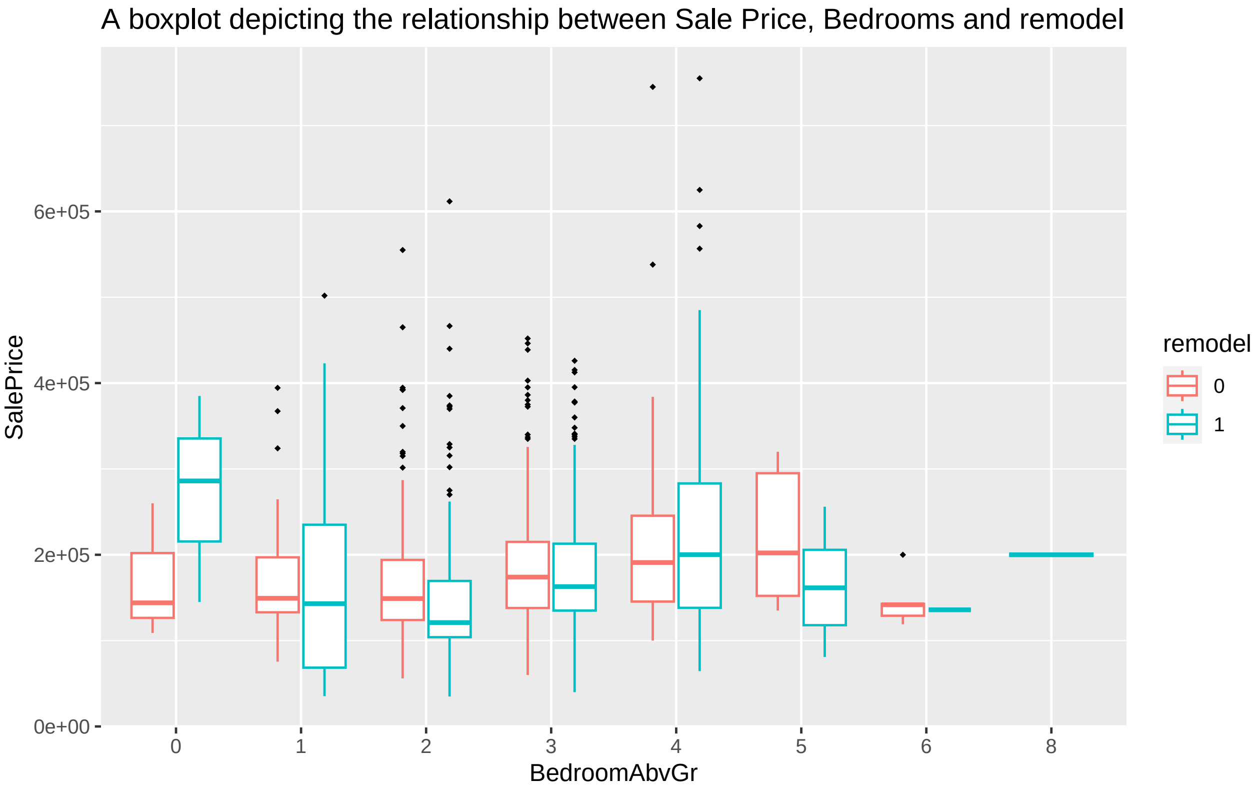


Figure 8: A boxplot depicting the relationship between Sale Price and Bedrooms above grade with a hue showing whether the home was remodeled or not

```
ggplot(Q4_df, aes(x = YearBuilt, y = SalePrice,
    color = remodel)) + geom_point(size = 2, shape = 19) +
    labs(title = "Scatter plot depicting the relationship between Sale Price, YearBuilt and remodel")
```

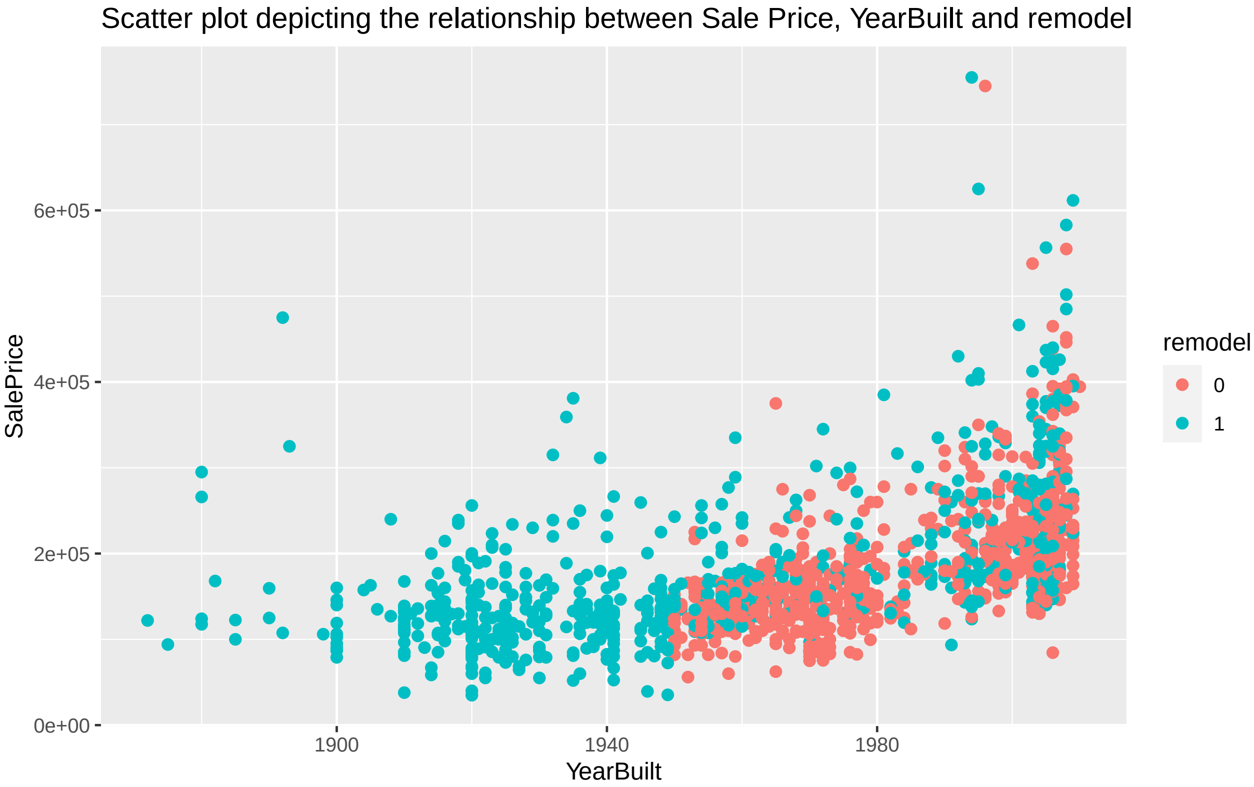


Figure 9: A scatter plot depicting the relationship between Sale Price and YearBuilt with a hue showing whether the home was remodeled or not

8) **OPTIONAL: PICK 2 OF 4** Discuss briefly any ethical concerns like residual disclosure that might arise from the use of your data set, possibly in combination with some additional data outside your dataset.

When working with data it is always crucial to identify potential ethics concerns that could arise from the data's availability to the public. The dataset I used for this project is the housing data from Ames between the years 1872 to 2010, the data is very detailed including columns for year and month sold, Neighborhood detailing the physical location within Ames city limits, Proximity to various conditions(i.e. Artery Adjacent to arterial street, Feedr Adjacent to feeder street, Norm Normal, RRNn Within 200' of North-South Railroad), etc. Firstly this could be a major ethics concern since having all these details about a home combined with a dataset including home purchases in the same or other cities or areas could potentially lead to the ability to identify people and follow them to where they move. Secondly this dataset also contains data on the types of road access, LotFrontage which captures the linear feet of street connected to property(0 if none), Street which captures type of road access to property, and Alley which captures the type of alley access to the property(NA if none). Having detailed information on the types of access to a property combined with the area and price of a home could lead to a homeowner becoming a target for a potential robbery if the home has a alley access, which could infer an easier getaway and multiple modes of entry, and is very large and pricey meaning the owner is most likely wealthy and would be more likely to be storing expensive items in the home. Lastly under the condition that you combine this dataset with a dataset containing all the people who work in the cities using details about the city like population and density this could potentially lead to people being able to identify where specific individuals live.