Effect of Appraised Land Value, Appraised Improvements Value and Neighbourhood District on Property Sales Price

Submitted to:
Professor Joslin Goh

Report Prepared By:
Sung Yu
Chen Qian
Mandy Yuan Yuan Gu
Quinten Carfagnini

July 19th, 2016

Table of Contents

Objective Summary

This report summarizes the relationship between appraised land value, appraised improvements value and neighbourhood district (3 explanatory variables) and the property sales price (response variable) in the city of Tampa, Florida between May 2008 and June 2009. The purpose of this report is to provide useful information for real estate appraisers, tax assessors, real estate investors and homebuyers.

To create the statistical model, we removed appraised land value as an explanatory variable and took the ln transformation of the property sales price and appraised improvements value. Our model passed all tests of model adequacy. Afterwards, we used automatic model selection to fit our model.

We found that there is a positive correlation between appraised improvements value and property sales price. Furthermore, the property sales price is also affected by the neighbourhood district: however, this correlation is stronger in certain neighbourhood districts than others.

Model Selection

Data Transformation

First, we took the ln transformation of property sales price, appraised improvements value and appraised land value to stabilize the variance. We will call this model “Model 1”. For all subsequent models and representations, *y* represents the property sales price.

$$y = \beta_1 x_{IMP} + \beta_2 x_{LAND} + \beta_3 x_{AVILA} + \beta_4 x_{CARROLLWOOD} + \beta_5 x_{CHEVAL} + \beta_6 x_{DAVISISLES} + \beta_7 x_{HUNTERSGREEN} + \beta_8 x_{HYDEPARK} + \beta_9 x_{TAMPAPALMS} + \beta_{10} x_{TOWNnCNTRY} + \varepsilon$$

Multicollinearity Issues with Model 1

Using the *vif()* function in R, we found the VIF between neighbourhood district and the other explanatory variables to be 11.27 and the VIF between appraised land value and the other explanatory variables to be 9.97. This is strong evidence of multicollinearity.

Removing an Explanatory Variable (Appraised Land Value)

Figure 7 - Scatterplot Matrix of Explanatory Variables (see Appendix)

One remedy of multicollinearity is to remove an explanatory variable. By comparing the relationship between the explanatory variables, we found that appraised land value and neighbourhood district has the strongest correlation.

We decided to remove appraised land value as properties in the same neighbourhood district usually have very similar types of land. As a result, the neighbourhood district would be able to give a good estimate of the appraised land value. However, the reverse is not necessarily true as two or more neighbourhood districts can have the same appraised land value.

Outliers

Figure 5 - Box Plot of Final Model (see Appendix)

We identified potential outliers using the *outlierTest* function from *library(car)* and plotted them against a box plot. After a close examination of the potential outliers, we did not believe that they were abnormal since there are other factors (such as the buyer’s strength of appeal and the seller’s urgency to sell) which affects property sales price. Given the absence of an expert opinion, we decided not to eliminate any of the potential outliers.

Leverages and Influential Measures

Figure 6 - Cook's Influence Measure vs Leverage (see Appendix)

We identified the potential leverage and influential points using Cook's Distance. After a close examination of the potential leverage and influential points, we decided not to remove them as our model is able to pass all tests of model adequacy while containing these points.

Removing the Base Case Dummy Variable

We decided to remove the default base case dummy variable (neighbourhood district Avila) for our model to make model results easier to interpret.

Automatic Model Selection Results

The results from stepwise model selection, forward model selection and backwards model selection all converge to the same model. We will call this model "Model 2".

Model 2

$$y = \beta_1 x_{AVILA} + \beta_2 x_{CARROLLWOOD} + \beta_3 x_{CHEVAL} + \beta_4 x_{DAVISISLES} + \beta_5 x_{HUNTERSGREEN} + \beta_6 x_{HYDEPARK} + \beta_7 x_{TAMPAPALMS} + \beta_8 x_{TOWNnCNTRY} + \beta_9 x_{IMP:AVILA} + \beta_{10} x_{IMP:CARROLLWOOD} + \beta_{11} x_{IMP:CHEVAL} + \beta_{12} x_{IMP:DAVISISLES} + \beta_{13} x_{IMP:HUNTERSGREEN} + \beta_{14} x_{IMP:HYDEPARK} + \beta_{15} x_{IMP:TAMPAPALMS} + \beta_{16} x_{IMP:TOWNnCNTRY} + \varepsilon$$

Manual Model Selection: AIC

After fitting our model, we used AIC to compare Model 2 against the model with no interaction effects.

Model with no interaction effects:

$$y = 0.67 x_{IMP} + 2.83 x_{AVILA} + 2.26 x_{CARROLLWOOD} + 2.42 x_{CHEVAL} + 3.12 x_{DAVISISLES} + 2.22 x_{HUNTERSGREEN} + 2.79 x_{HYDEPARK} + 2.35 x_{TAMPAPALMS} + 1.95 x_{TOWNnCNTRY}$$

AIC: 46.08975

Model 2 (with fitted values):

$$y = 1.97 x_{AVILA} + 1.82 x_{CARROLLWOOD} + 1.27 x_{CHEVAL} + 4.53 x_{DAVISISLES} + 0.58 x_{HUNTERSGREEN} + 2.89 x_{HYDEPARK} + 0.92 x_{TAMPAPALMS} + 1.53 x_{TOWNnCNTRY} + 0.8 x_{IMP:AVILA} + 0.75 x_{IMP:CARROLLWOOD} + 0.88 x_{IMP:CHEVAL} + 0.39 x_{IMP:DAVISISLES} + 0.98 x_{IMP:HUNTERSGREEN} + 0.65 x_{IMP:HYDEPARK} + 0.94 x_{IMP:TAMPAPALMS} + 0.76 x_{IMP:TOWNnCNTRY}$$

AIC: -64.89734

The AIC shows that Model 2 is the better model.

Selecting Model 2 as our Final Model

The adjusted R-squared of Model 2 is 0.9987. Moreover, we believe that Model 2 does not have an issue with overfitting as we have sufficient degrees of freedom (334). As a result, we selected Model 2 as our final model.

Model Diagnostics

All tests of model adequacy are performed on Model 2. See Appendix for the figures below.

Figure 1 – Standardized Residuals vs Fitted Values

Figure 2 - Histogram of Standardized Residuals

Figure 3 – Standardized Residuals vs Explanatory Variable (Appraised Improvements Value)

Figure 4 - QQ-Plot of Residuals

The above residual plots and tests of model adequacy show no obvious violations of assumptions. The standardized residuals appear to be normally distributed and there are no noticeable patterns between the standardized residuals and the fitted values, nor is there any noticeable patterns between the standardized residuals and appraised improvements value.

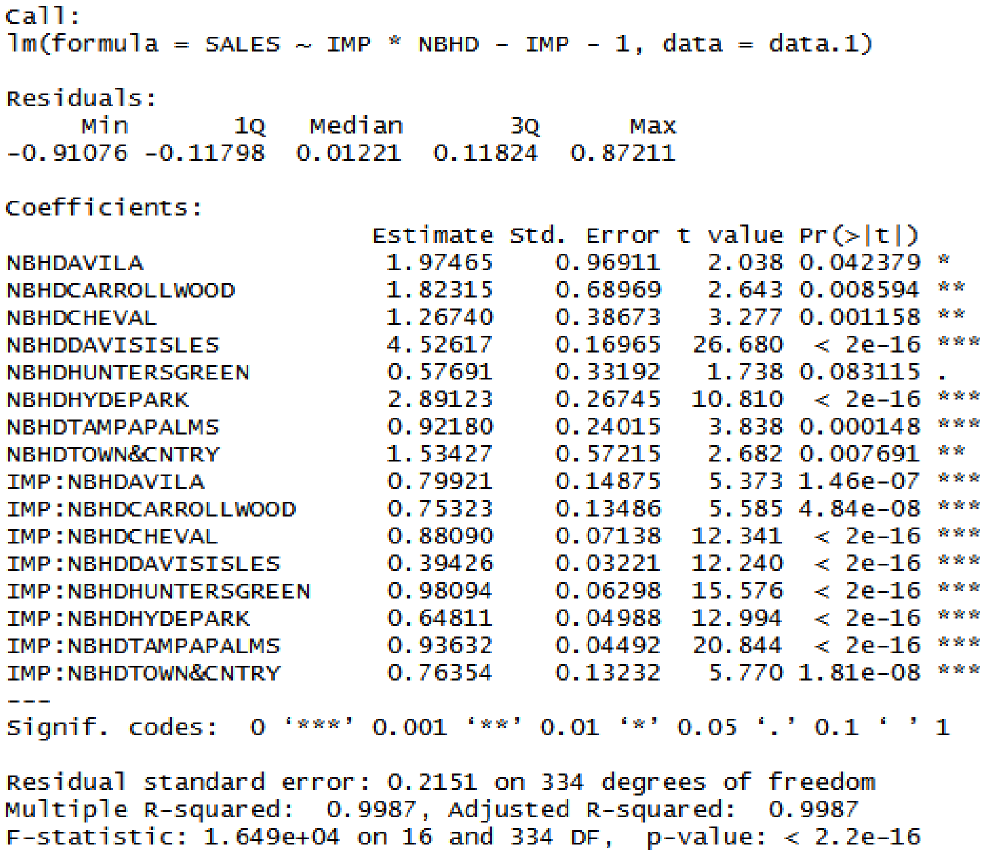
Thus, Model 2 passes the tests of model adequacy.

Conclusion

Our final model is:

$$y = 1.97x_{AVILA} + 1.82x_{CARROLLWOOD} + 1.27x_{CHEVAL} + 4.53x_{DAVISISLES} + 0.58x_{HUNTERSGREEN} + 2.89x_{HYDEPARK} + 0.92x_{TAMPAPALMS} + 1.53x_{TOWNnCNTRY} + 0.8x_{IMP:AVILA} + 0.75x_{IMP:CARROLLWOOD} + 0.88x_{IMP:CHEVAL} + 0.39x_{IMP:DAVISISLES} + 0.98x_{IMP:HUNTERSGREEN} + 0.65x_{IMP:HYDEPARK} + 0.94x_{IMP:TAMPAPALMS} + 0.76x_{IMP:TOWNnCNTRY}$$

Here is the summary for the model. Please note that because we took the ln transformation of the property sales price and appraised improvements value, we need to take *e* to the power of the coefficient for the actual interpretation.



```
Call:
lm(formula = SALES ~ IMP * NBHD - IMP - 1, data = data.1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91076 -0.11798  0.01221  0.11824  0.87211

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
NBHDAVILA                1.97465    0.96911   2.038 0.042379 *
NBHDCARROLLWOOD          1.82315    0.68969   2.643 0.008594 **
NBHDCHEVAL               1.26740    0.38673   3.277 0.001158 **
NBHDDAVISISLES           4.52617    0.16965  26.680  < 2e-16 ***
NBHDHUNTERSGREEN         0.57691    0.33192   1.738 0.083115 .
NBHDHYDEPARK             2.89123    0.26745  10.810  < 2e-16 ***
NBHDTAMPAPALMS           0.92180    0.24015   3.838 0.000148 ***
NBHDTOWN&CNTRY           1.53427    0.57215   2.682 0.007691 **
IMP:NBHDAVILA            0.79921    0.14875   5.373 1.46e-07 ***
IMP:NBHDCARROLLWOOD      0.75323    0.13486   5.585 4.84e-08 ***
IMP:NBHDCHEVAL           0.88090    0.07138  12.341  < 2e-16 ***
IMP:NBHDDAVISISLES       0.39426    0.03221  12.240  < 2e-16 ***
IMP:NBHDHUNTERSGREEN     0.98094    0.06298  15.576  < 2e-16 ***
IMP:NBHDHYDEPARK         0.64811    0.04988  12.994  < 2e-16 ***
IMP:NBHDTAMPAPALMS       0.93632    0.04492  20.844  < 2e-16 ***
IMP:NBHDTOWN&CNTRY       0.76354    0.13232   5.770 1.81e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2151 on 334 degrees of freedom
Multiple R-squared:  0.9987, Adjusted R-squared:  0.9987
F-statistic: 1.649e+04 on 16 and 334 DF,  p-value: < 2.2e-16
```

The model suggests that there is a positive correlation between property sales price and appraised improvements value. There is also an interaction effect between neighbourhood district and appraised improvements value. This suggests that the rate at which improvements affects sales price is different in each neighbourhood district. On the other hand, appraised land value and property sales price are not linearly related. The influence of appraised land value is a constant amount which varies by neighbourhood district.

Reflection

Appraisal Criteria

As shown in the summary of the model found in the “Conclusions” section, the prediction equation relating property sales price to the appraised improvements value is different for each neighbourhood. This suggests that the appraisers do not use the same appraisal criteria for each neighbourhood.

Predictive Power of the Model: A Warning

Although our model passes the tests of model adequacy, there is no assurance that our model would be a successful predictor of future property sales price. There are many other factors which affect the sales price. This includes, but is not limited to: the economy, housing market and interest rates. As these factors are constantly changing, we would not be able to accurately predict future sales price with a model based on data from May 2008-June 2009.

Improving our Model

We handled the multicollinearity issue by removing an explanatory variable. However, there are many other methods of treating multicollinearity such as partial least squares regression and principal component analysis. There is a possibility that using either of these methods as opposed to removing an explanatory variable could improve our model.

Furthermore, an expert opinion can help improve our model. An industry expert can make a more qualified judgement on whether or not we should remove any of the potential outliers, high leverage or influential points.

Appendix

Figures

Figure 1 – Standardized Residuals vs Fitted Values for Model 2

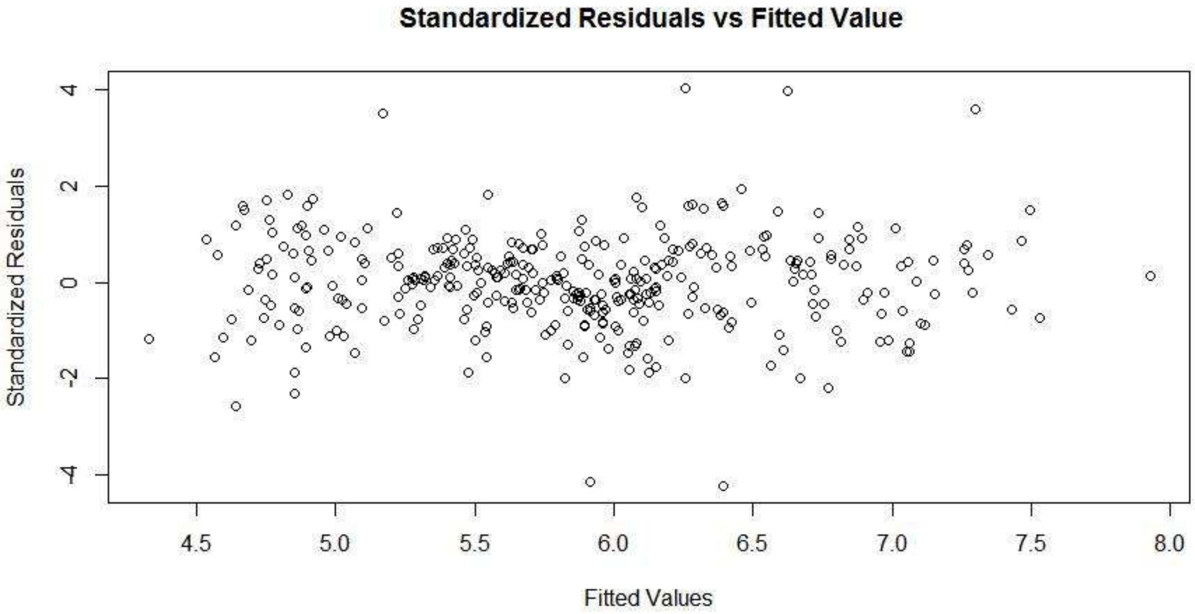


Figure 2 - Histogram of Standardized Residuals for Model 2

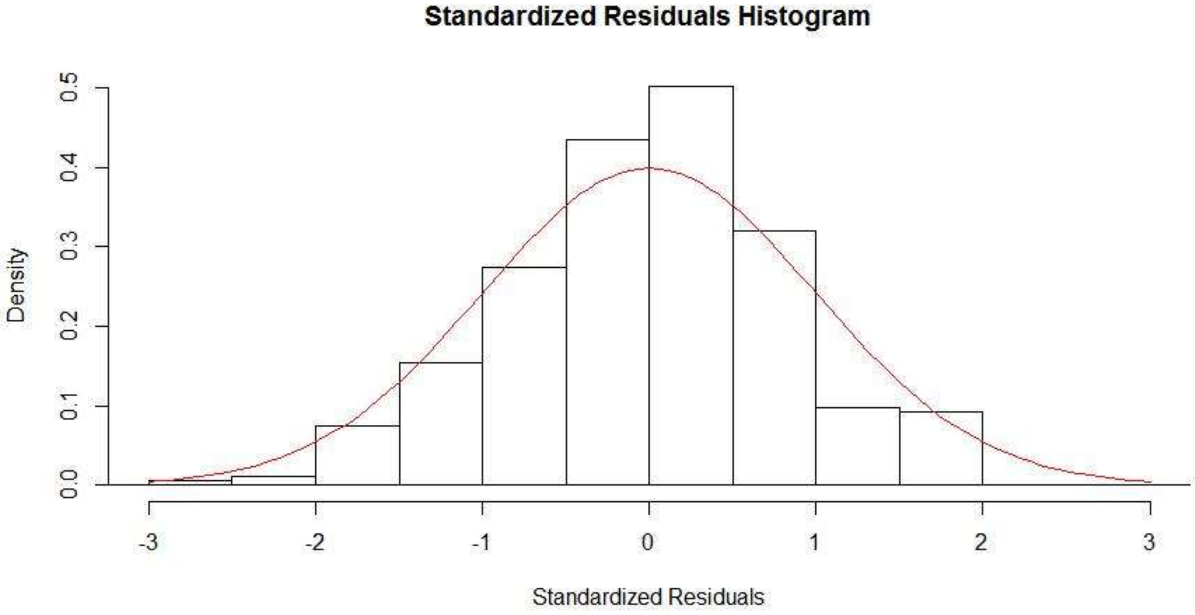


Figure 3 – Standardized Residuals vs Explanatory Variable (Appraised Improvements Value)

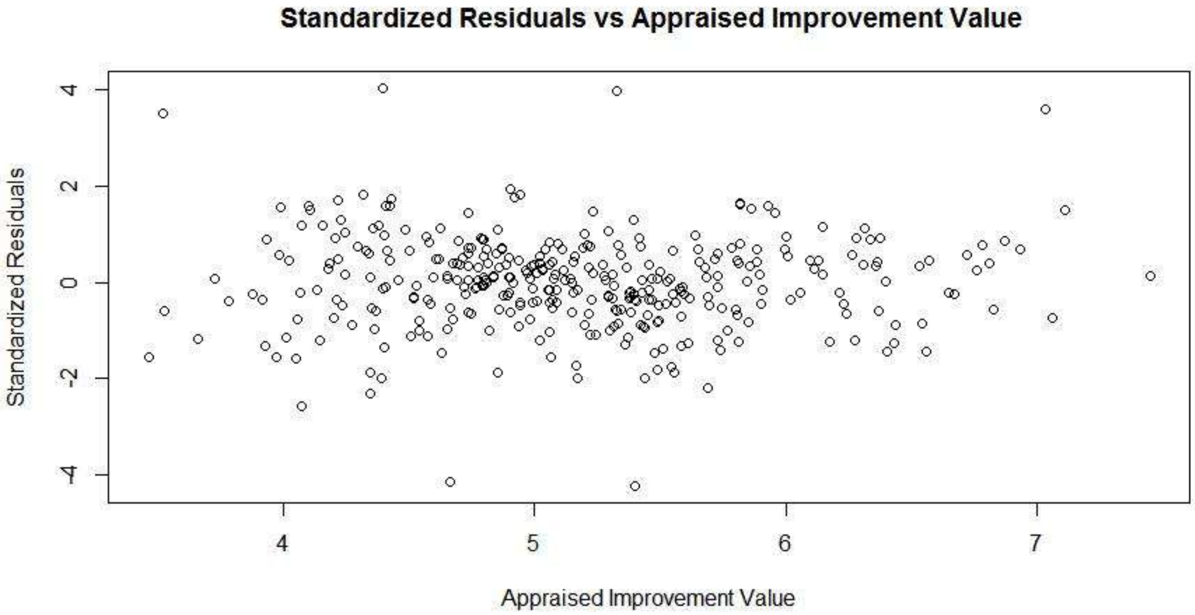


Figure 4 - QQ-Plot of Residuals for Model 2

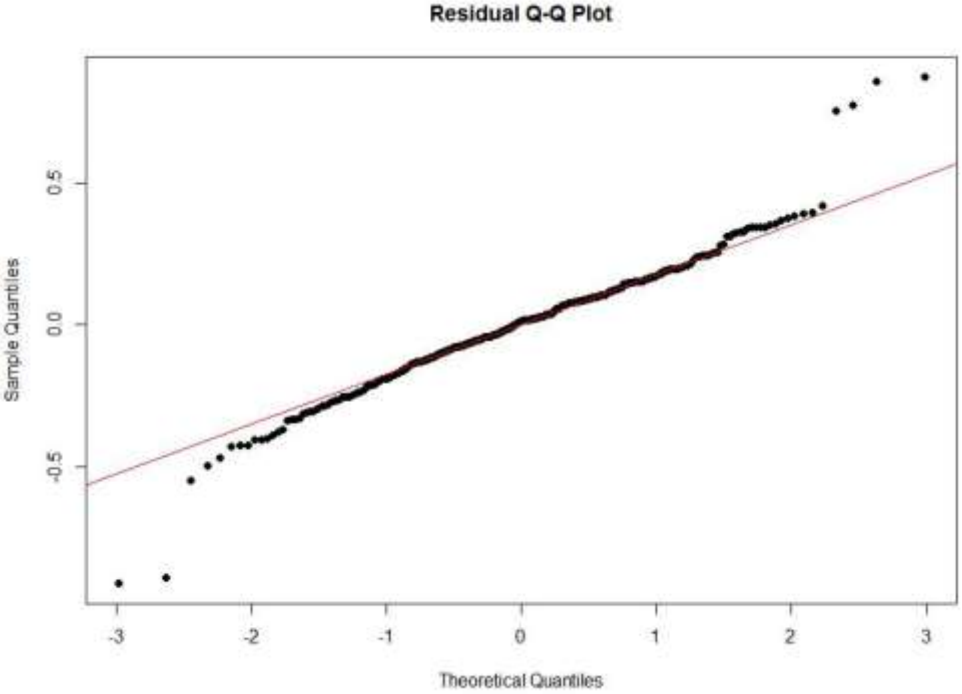


Figure 5 - Box Plot for Model 2

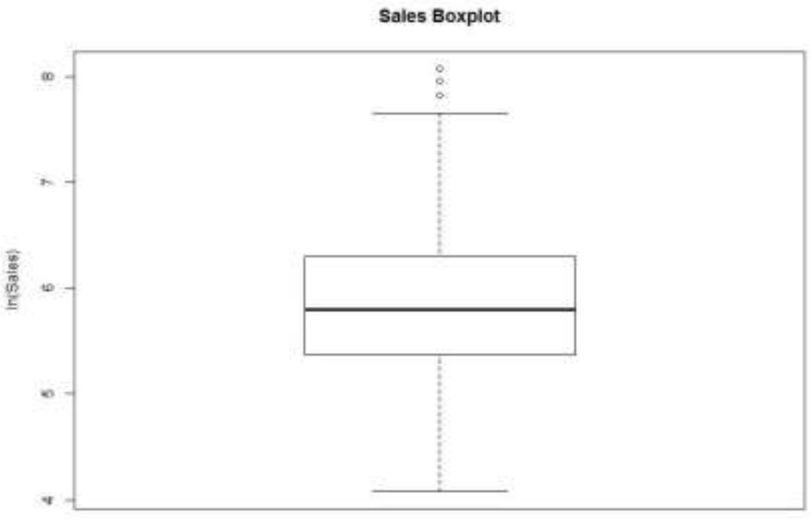


Figure 6 - Cook’s Influence Measure vs Leverage

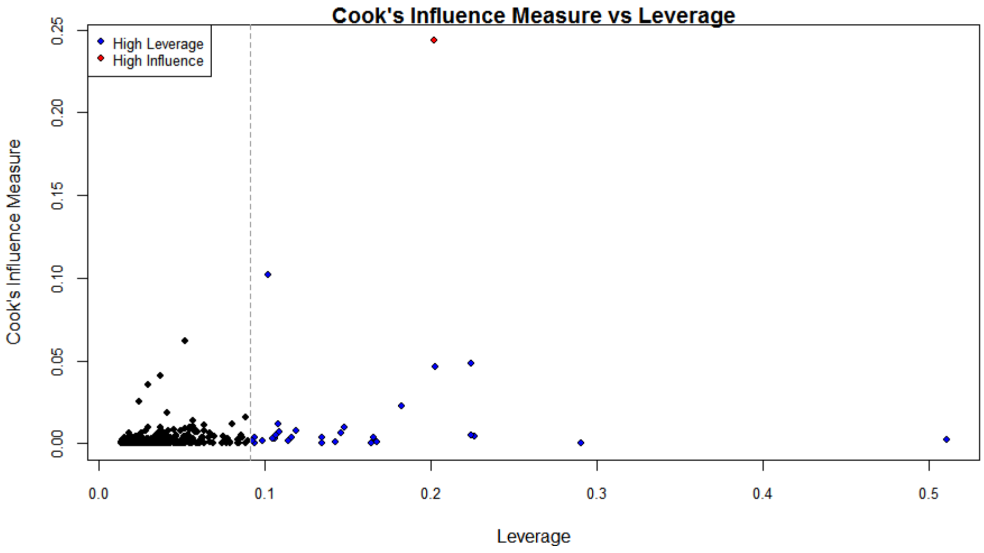
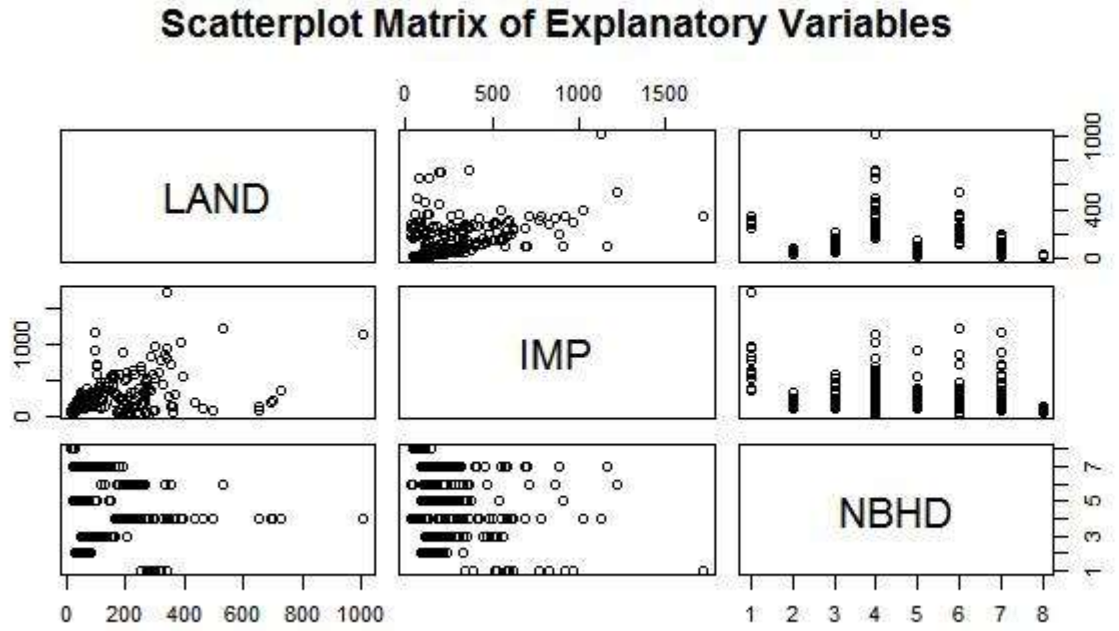
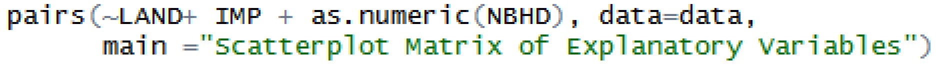


Figure 7 - Scatterplot Matrix of Explanatory Variables



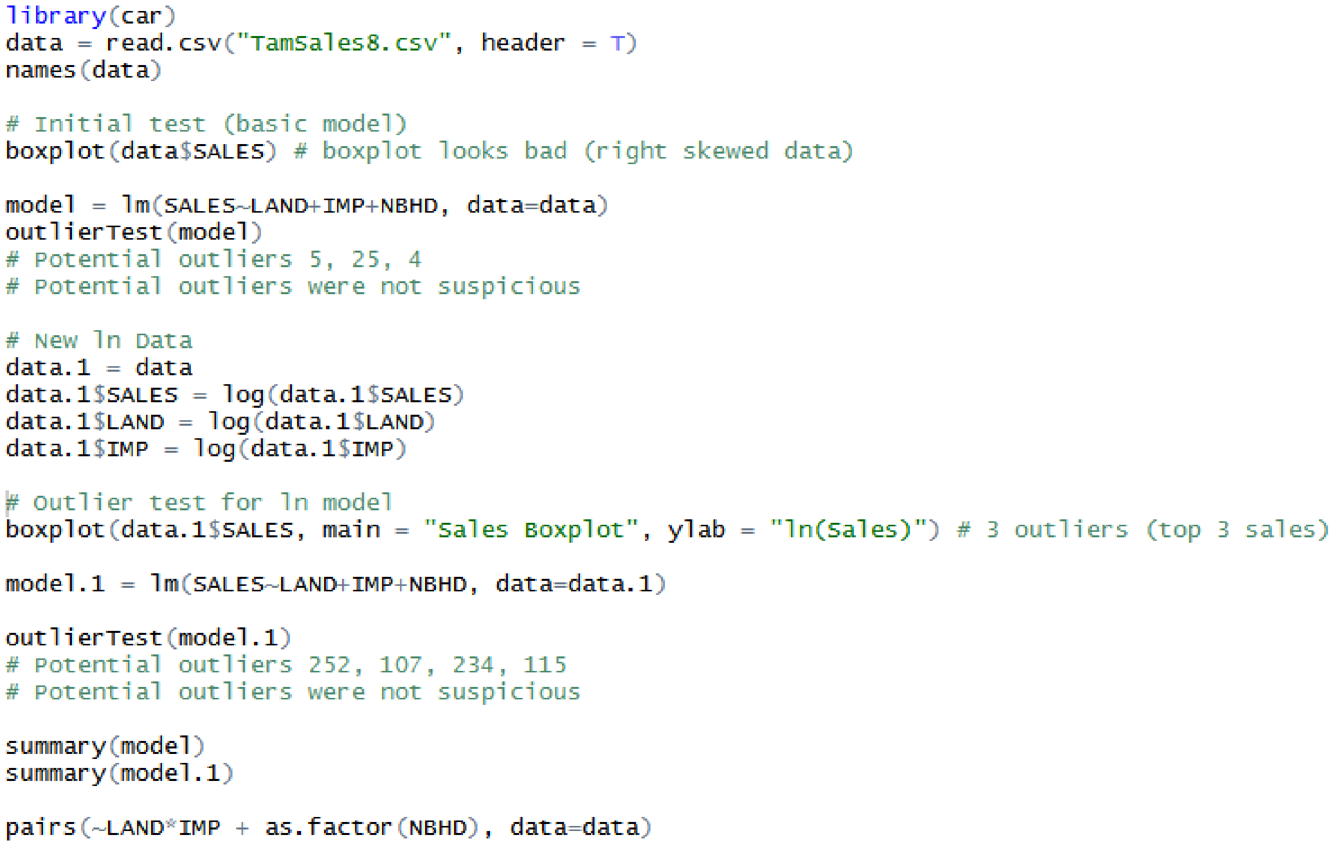
R Code

Matrix Scatterplot for LAND, IMP and NBHD



```
pairs(~LAND+ IMP + as.numeric(NBHD), data=data,
      main ="Scatterplot Matrix of Explanatory Variables")
```

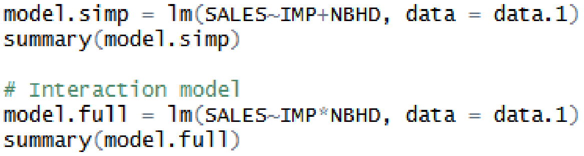
Outliers and Data Transformation



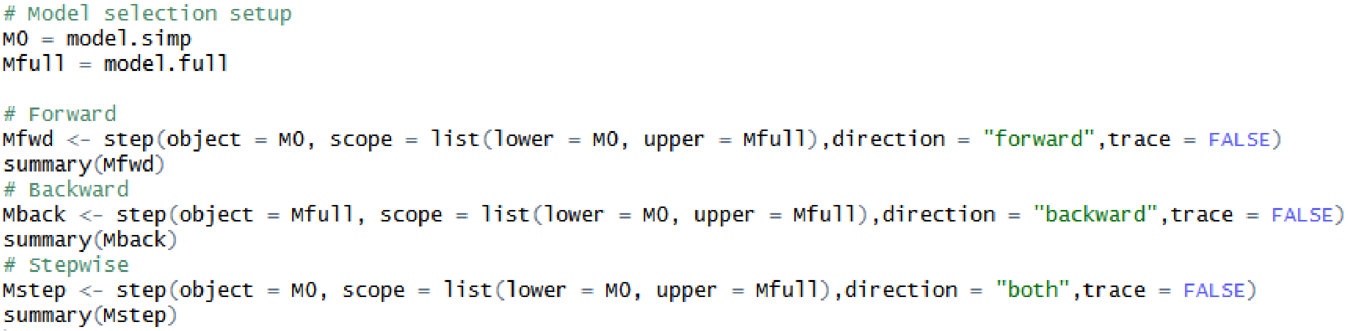
```
library(car)
data = read.csv("TamSales8.csv", header = T)
names(data)

# Initial test (basic model)
boxplot(data$SALES) # boxplot looks bad (right skewed data)

model = lm(SALES~LAND+IMP+NBHD, data=data)
outlierTest(model)
# Potential outliers 5, 25, 4
# Potential outliers were not suspicious

# New ln Data
data.1 = data
data.1$SALES = log(data.1$SALES)
data.1$LAND = log(data.1$LAND)
data.1$IMP = log(data.1$IMP)

# Outlier test for ln model
boxplot(data.1$SALES, main = "Sales Boxplot", ylab = "ln(Sales)") # 3 outliers (top 3 sales)

model.1 = lm(SALES~LAND+IMP+NBHD, data=data.1)

outlierTest(model.1)
# Potential outliers 252, 107, 234, 115
# Potential outliers were not suspicious

summary(model)
summary(model.1)

pairs(~LAND*IMP + as.factor(NBHD), data=data)
```

Automatic Model Selection

# Model.simp is the model with appraised land values removed and no interaction effects.

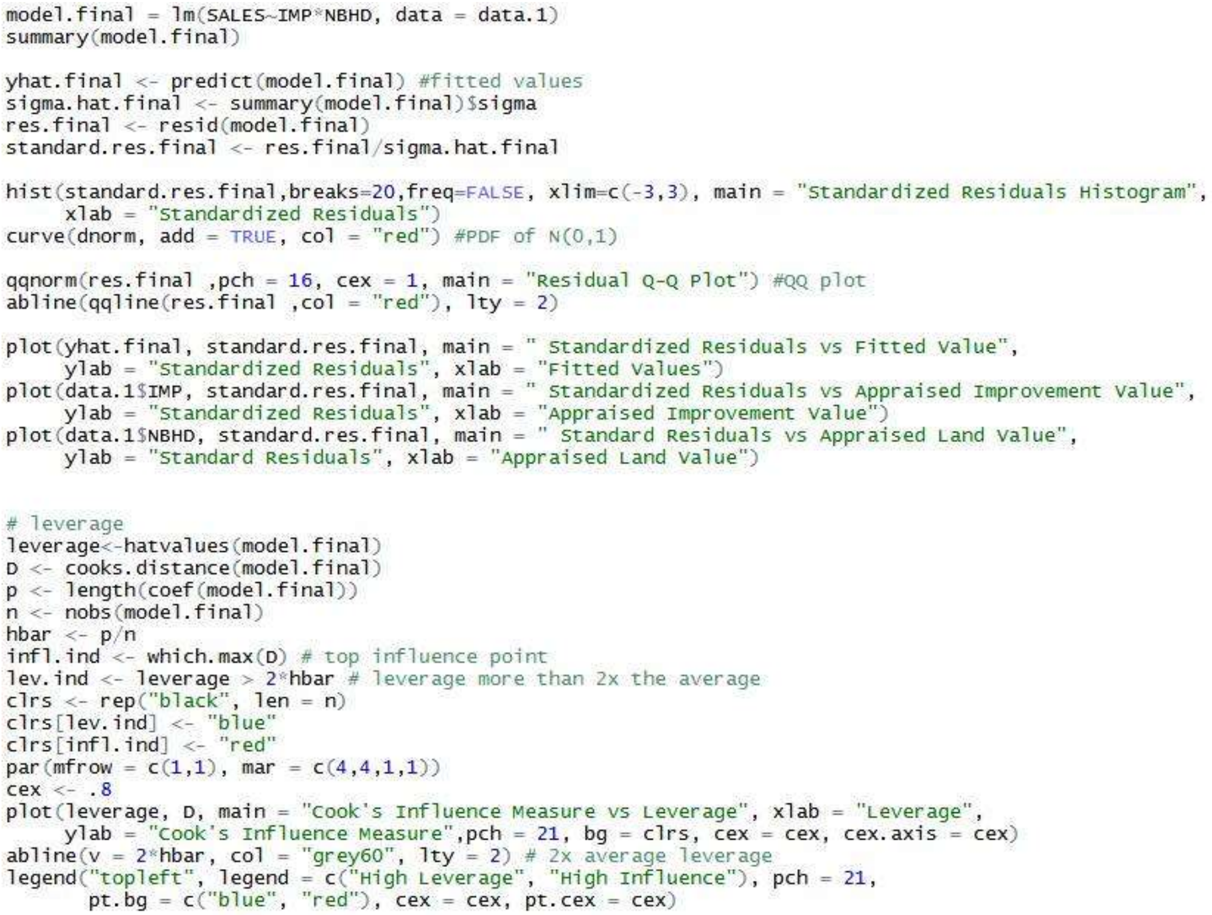


```
model.simp = lm(SALES~IMP+NBHD, data = data.1)
summary(model.simp)

# Interaction model
model.full = lm(SALES~IMP*NBHD, data = data.1)
summary(model.full)
```



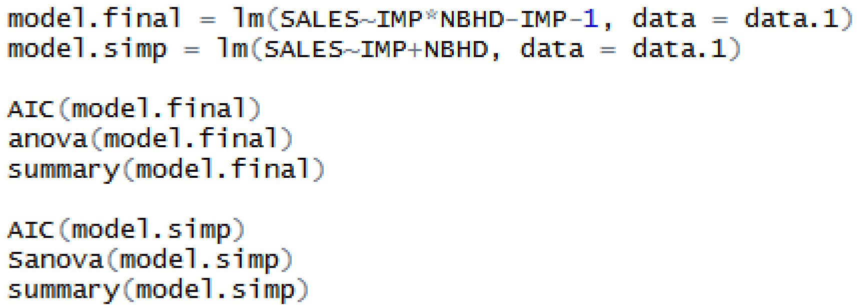
```
# Model selection setup
M0 = model.simp
Mfull = model.full

# Forward
Mfwd <- step(object = M0, scope = list(lower = M0, upper = Mfull),direction = "forward",trace = FALSE)
summary(Mfwd)
# Backward
Mback <- step(object = Mfull, scope = list(lower = M0, upper = Mfull),direction = "backward",trace = FALSE)
summary(Mback)
# Stepwise
Mstep <- step(object = M0, scope = list(lower = M0, upper = Mfull),direction = "both",trace = FALSE)
summary(Mstep)
```

Residual Checking and Influential Points



```
model.final = lm(SALES~IMP*NBHD, data = data.1)
summary(model.final)

yhat.final <- predict(model.final) #fitted values
sigma.hat.final <- summary(model.final)$sigma
res.final <- resid(model.final)
standard.res.final <- res.final/sigma.hat.final

hist(standard.res.final,breaks=20,freq=FALSE, xlim=c(-3,3), main = "Standardized Residuals Histogram",
     xlab = "Standardized Residuals")
curve(dnorm, add = TRUE, col = "red") #PDF of N(0,1)

qqnorm(res.final ,pch = 16, cex = 1, main = "Residual Q-Q Plot") #QQ plot
abline(qqline(res.final ,col = "red"), lty = 2)

plot(yhat.final, standard.res.final, main = " Standardized Residuals vs Fitted Value",
     ylab = "Standardized Residuals", xlab = "Fitted Values")
plot(data.1$IMP, standard.res.final, main = " Standardized Residuals vs Appraised Improvement Value",
     ylab = "Standardized Residuals", xlab = "Appraised Improvement Value")
plot(data.1$NBHD, standard.res.final, main = " Standard Residuals vs Appraised Land Value",
     ylab = "Standard Residuals", xlab = "Appraised Land Value")


# leverage
leverage<-hatvalues(model.final)
D <- cooks.distance(model.final)
p <- length(coef(model.final))
n <- nobs(model.final)
hbar <- p/n
infl.ind <- which.max(D) # top influence point
lev.ind <- leverage > 2*hbar # leverage more than 2x the average
clrs <- rep("black", len = n)
clrs[lev.ind] <- "blue"
clrs[infl.ind] <- "red"
par(mfrow = c(1,1), mar = c(4,4,1,1))
cex <- .8
plot(leverage, D, main = "Cook's Influence Measure vs Leverage", xlab = "Leverage",
     ylab = "Cook's Influence Measure",pch = 21, bg = clrs, cex = cex, cex.axis = cex)
abline(v = 2*hbar, col = "grey60", lty = 2) # 2x average leverage
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21,
       pt.bg = c("blue", "red"), cex = cex, pt.cex = cex)
```

Model Checking with AIC



```
model.final = lm(SALES~IMP*NBHD-IMP-1, data = data.1)
model.simp = lm(SALES~IMP+NBHD, data = data.1)

AIC(model.final)
anova(model.final)
summary(model.final)

AIC(model.simp)
Sanova(model.simp)
summary(model.simp)
```