Assignment #2
Jennifer M. Wanat

Introduction:

In this report our objective is to be able to provide estimates of home values for the typical home in Ames, Iowa. The data set was obtained from the Ames, Iowa Assessor's Office and assembled by Dr. Dean De Cock at Truman State University. In order to build a model to provide estimates of home values, an exploratory data analysis (EDA) of the data set was completed, which allowed for the selection of two of the most promising predictor variables.

The data documentation was reviewed to understand the types of variables collected. As the goal is to predict the typical home value, some conditions were dropped so that the data set represented the typical single-family home. Then an EDA was conducted to demonstrate why the two variables were selected for modeling purposes. Descriptive statistics was performed on the variables when possible. The variables were tabulated and examined for null entries or errors.

Next, the two predictor variables were each fitted in a simple linear regression model, and then combined into a multiple linear regression model. The model summaries and parameters were provided, and the goodness-of-fit was assessed. Finally, the three models were refit using the log transformation of sale price. The EDA and linear regression modeling were conducted with the R programming language.

Data:

The data set contains 82 variables measured from 2930 individual residential properties sold in Ames, IA from 2006 to 2010. Refer to the data documentation for description of the variables.

Sample Definition:

From the data set, the conditions listed in Figure 1 were dropped and not used for the sample population. This process of elimination is referenced as waterfall conditions, as the first condition dropped will result in a smaller sample size in which the subsequent condition to be dropped would be applied. This process continues until all waterfall conditions have been processed.

Figure 1: Waterfall conditions.

Variable	Drop Condition	Number of Properties Dropped
Building Type	Not equal to single-family detached	505
Sale Condition	Not equal to normal	423
Street	Not paved	6
Above Grade Living Area	Greater than 4,000 square feet	1

Lot Area	Greater than 100,000 square feet	3
Bedroom	No bedrooms	4
Full Bath	No full baths	1

The resulting sample population data set (a data frame called eligible.population) contained 82 variables from 1987 individual residential properties. All subsequent data quality checks and exploratory data analysis were conducted on this data set.

The waterfall conditions were selected to create a data set that represented typical single-family, detached homes in Ames, IA. The data documentation indicated that there were 5 observations from the original data set that were either outliers or unusual sales. The drop condition for observations greater than 4,000 square feet dropped these observations from the data set, if the prior waterfall conditions had not already done so.

Exploratory Data Analysis:

Two variables were selected from the original 82 variables of the data set for an exploratory data analysis. These variables were selected as a result of an initial EDA, and are thought to be the two most promising predictor variables for predicting sale price. The two variables selected were Gr Liv Area (the above grade (ground) living area in square feet) and Full Bath (the number of full bathrooms above grade). The EDA utilized boxplots, histograms and scatterplots, as appropriate.

As Gr Liv Area is a continuous variable, a basic descriptive statistical and quantile summary of the variable was conducted. The descriptive statistical summary in Figure 2 includes the minimum value (Min.), first quantile value (1st Qu.), median, mean, third quantile value (3rd Qu.), and maximum value (Max.) of the variable. The quantile summary in Figure 3 includes the variable value at the 0, 25, 50, 75 and 100% quantiles.

Figure 2: Descriptive statistics of GrLivArea variable.

Variable	Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
GrLivArea	334	1112	1445	1494	1760	3820

The statistics listed in Figure 2 are measures of central tendency. The mean or average is the sum of the measurements divided by the total number of measurements. The median is the middle value from the ordered set of measurements. The mean is affected more by outliers than the median. Quartiles (Q) divide the group of data into four equal parts. Q1 and Q3 are the first and third quartiles, and divide the data into the 25^{th} and 75^{th} percentiles, respectively. The median is also known as Q2 and is located at the 50^{th} percentile. Together, the data between Q1 and Q3 represent the middle 50% of the data. Minimum is the smallest value in the data set and maximum is the largest value in the data set.

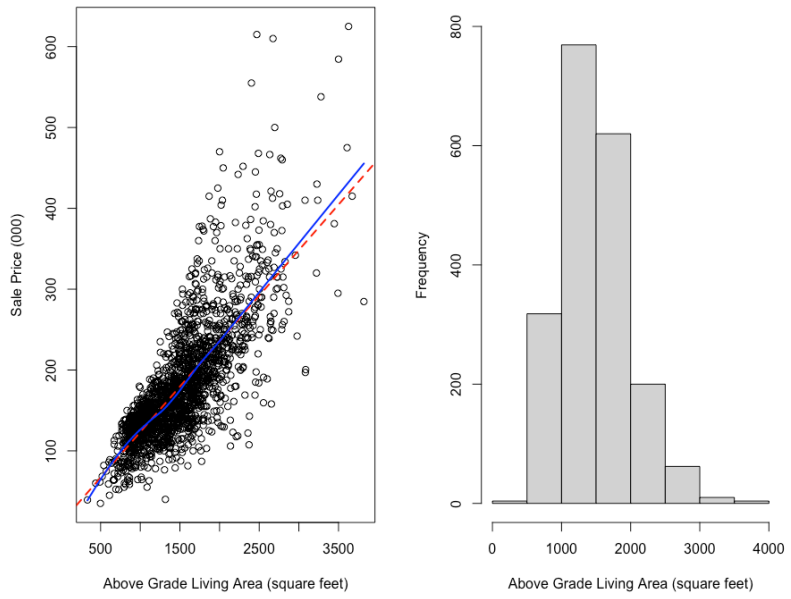
Figure 3: Quantile summary.

Variable	0%	25%	50%	75%	100%
GrLivArea	334	1112	1445	1759	3820

A scatterplot of sale price versus above grade living area is displayed in Figure 4a The loess smoother line and the regression line have a strong agreement between the two lines in the scatterplot. Both lines have a positive slope. A heteroscedastic, non-constant variance, is displayed as a widening cone of plotted values as sale price and above grade living area increases.

The histogram in Figure 4b indicates that most homes have less than 2000 square feet above grade living area, with the majority between 1000 and 1500 square feet.

Figure 4: (a) Scatterplot of sale price versus above grade living area. The dashed line was estimated using ordinary least squares or OLS. The solid line is a loess smoother. (b) Histogram of above grade living area.



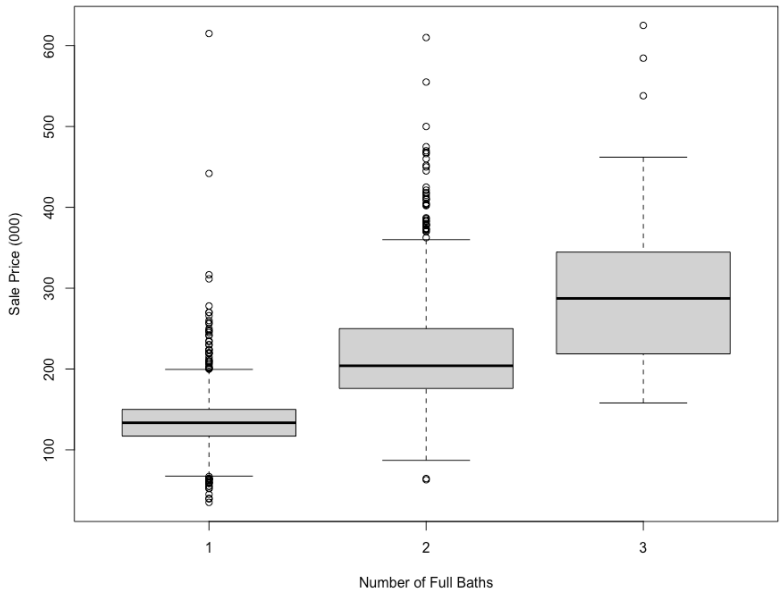
(a) (b)

As the Full Bath variable is a non-continuous (discrete) variable, a table summarizing the counts of the factor levels is displayed in Figure 5. The column NA indicates if there were any observations that contained null values for the variable. A NA entry could be due to an inadvertent missed entry, a typo, or the information was not specified. There were no NAs for the Full Bath variable.

Figure 5: Count of homes with the number of full baths.

1 Full bath	2 Full bath	3 Full bath	NA	Total homes
997	954	36	0	1987

Figure 6: Boxplot of sale price versus number of full baths.



A boxplot of the sale price as a function of the number of full baths is displayed in Figure 6. Generally, the sale price increases with the number of full baths. A basic descriptive statistical and quantile summary was conducted on the Full Bath variable after it was converted from a discrete to factor variable. Refer to Figures 7 and 8.

Figure 7: Descriptive statistics of sale price versus full bath variable.

Number of Full Baths	Minimum	Median	Mean	Maximum
1	35000	133500	135688	615000
2	63000	204000	219191	610000
3	158000	287350	303810	625000

Figure 8: Quantile summary of sale price versus full bath variable.

Number of Full Baths	0%	25%	50%	75%	100%
1	35000	117000	133500	150000	615000
2	63000	176000	204000	250000	610000
3	158000	221875	287350	342316	625000

Simple Linear Regression Models:

Each of the two predictor variables were independently fit in simple linear regression models with sale price. The model summaries and parameters are provided and discussed. The goodness-

of-fit of each model was assessed with two diagnostic plots, a Q-Q plot and a plot of the model residuals versus the predictor variable.

In a Q-Q plot, a set of equally spaced quantiles is calculated based upon the sample size of the data set. The observed and ordered (from lowest to highest value) data set are plotted against the calculated quantiles. The data points, if normally distributed, will plot approximately along a straight line drawn through the first and third quantiles. Variations from a normal distribution will be detected by data points that drift away from the theoretical quantile line.

The residuals, or the estimated errors, are calculated by subtracting the expected (or predicted) value of Y obtained from the linear equation for the linear regression model from the observed value of Y. The residuals are plotted against the predictor variable. Residuals should be independent, normally distributed and have a constant variance. Residuals should also have a mean of zero.

A subset of the sample population data set was created for the regression models. Approximately 70% of the data set was used in the model training data set (a data frame called train.df).

Model #1 Gr Liv Area:

Model #1 is a simple linear regression model of sale price as a function of above grade living area.

Figure 9: Residuals for Model #1 Gr Liv Area

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-170335	-24842	-1402	20265	326177

Figure 10: Coefficients for Model #1 Gr Liv Area

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	11886.769	3968.610	2.995	0.00279
GrLivArea	112.120	2.518	44.521	< 2e-16

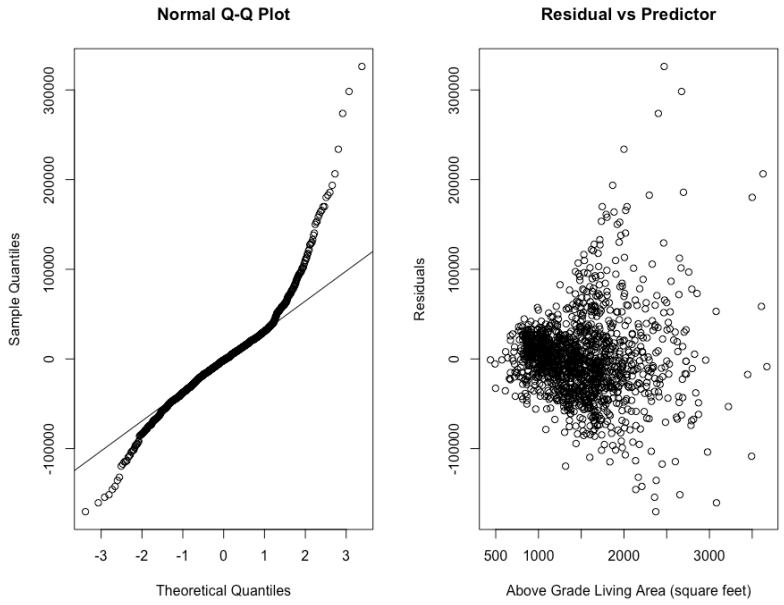
Figure 11: Model #1 summary

Residual standard error	45860 on 1402 degrees of freedom
Multiple R-squared	0.5857
Adjusted R-squared	0.5854
F-statistic	1982 on 1 and 1402 DF
p-value	< 2.2e-16

From the model, sale price is expected to increase by $112 for each square foot increase in above grade living area. The residual standard error is 45860. If model #1 was used to predict the sale price from above grade living area square footage, we can expect to be accurate to within approximately ±91720 dollars at a 95% confidence level. The multiple R-squared value is 0.5857, which indicates that 58.5% of the variation in sale price (about its mean) can be explained by the variable above grade living area. The linear association between sale price and

above grade living area is statistically significant at the 5% significance level, based upon the results of the hypothesis tests for the model coefficients.

Figure 12: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #1. (b) Residual versus predictor plot of model #1.



(a) (b)

The Q-Q plot of model #1 indicates a departure of values from a normal distribution for lower and higher sale priced homes. The residual versus predictor plot of model #1 demonstrates a heteroscedastic, non-constant variance, displayed as a widening cone of plotted values as the above grade living area increases.

Model #2 Full Bath:

Model #2 is a simple linear regression model of sale price as a function of the number of full baths.

Figure 13: Residuals for Model #2 Full Bath

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-155554	-31665	-5665	19335	478335

Figure 14: Coefficients for Model #2 Full Bath

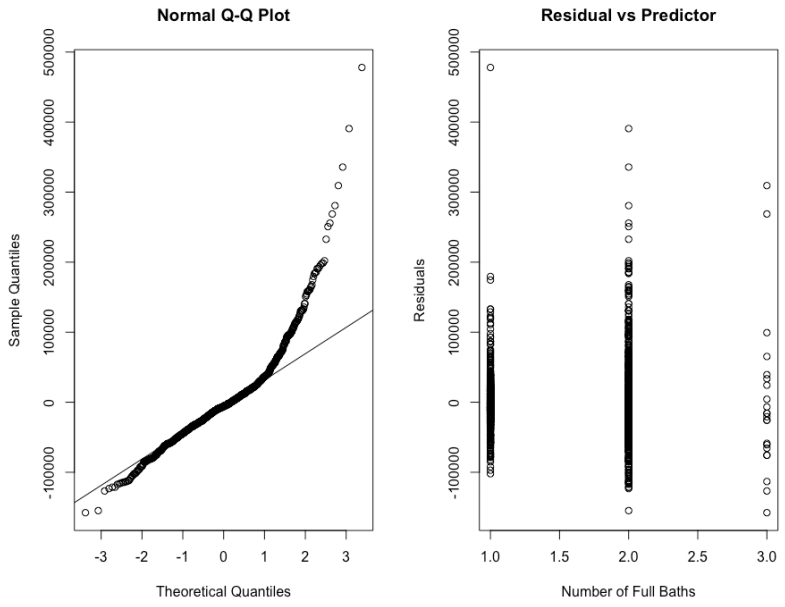
	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	53275	4525	11.77	< 2e-16
Full Bath	83389	2812	29.66	< 2e-16

Figure 15: Model #2 summary

Residual standard error	55850 on 1402 degrees of freedom
Multiple R-squared	0.3855
Adjusted R-squared	0.385
F-statistic	879.4 on 1 and 1402 DF
p-value	< 2.2e-16

From the model, sale price is expected to increase by $83389 for each full bath. The residual standard error is 55850. If model #2 was used to predict the sale price from above grade living area square footage, we can expect to be accurate to within approximately ±111700 dollars at a 95% confidence level. The multiple R-squared value is 0.3855, which indicates that 38.5% of the variation in sale price (about its mean) can be explained by the variable of number of full baths. The linear association between sale price and number of full baths is statistically significant at the 5% significance level, based upon the results of the hypothesis tests for the model coefficients.

Figure 16: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #2. (b) Residual versus predictor plot of model #2.



(a) (b)

The Q-Q plot of model #2 indicates a departure of values from a normal distribution for higher sale priced homes. The residual versus predictor plot of model #2 demonstrates a heteroscedastic, non-constant variance, displayed as a widening column of plotted values as the number of full baths increase.

Multiple Linear Regression Model – Model #3:

Model #3 is a multiple linear regression model of sale price as a function of the above grade living area and the number of full baths.

Figure 17: Residuals for Model #3

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-167380	-23737	-1592	18779	358662

Figure 18: Coefficients for Model #3

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	26108.551	4283.531	6.095	1.41e-9
GrLivArea	93.210	3.307	28.187	< 2e-16
Factor(Full Bath)2	27677.072	3083.411	8.976	< 2e-16
Factor(Full Bath)3	39688.086	10847.194	3.659	0.000263

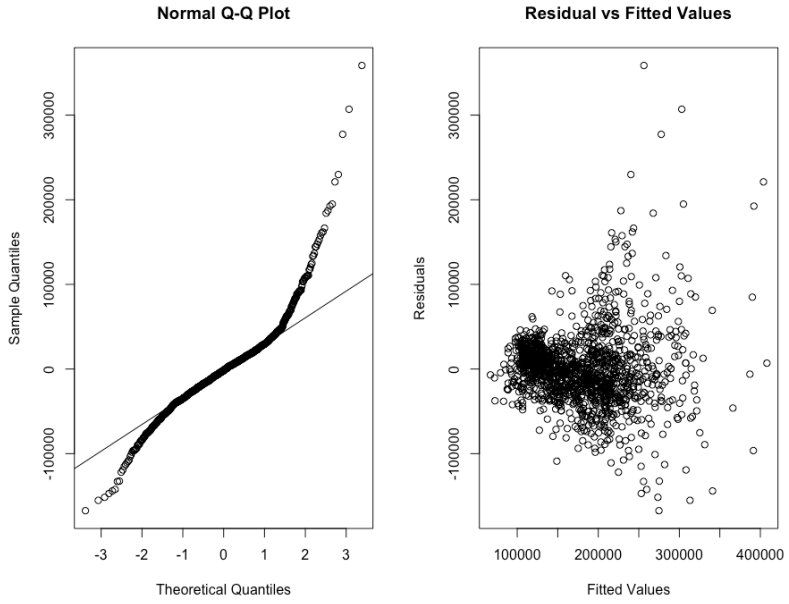
Figure 19: Model #3 summary

Residual standard error	44620 on 1400 degrees of freedom
Multiple R-squared	0.6083
Adjusted R-squared	0.6075
F-statistic	724.7 on 3 and 1400 DF
p-value	< 2.2e-16

From the model, sale price is expected to increase by $93.21 for each square foot increase in above grade living area when the number of full baths is held constant. The sale price is expected to increase by $27677 for a second full bath and increase by $39688 for a third full bath when the above grade living area remains constant. The residual standard error is 44620. If model #3 was used to predict the sale price from above grade living area square footage and number of full baths, we can expect to be accurate to within approximately ±59240 dollars at a 95% confidence level. The multiple R-squared value is 0.6083, which indicates that 60.8% of the variation in sale price (about its mean) can be explained by a multiple linear regression association between the above grade living area and the number of full baths. The linear association between sale price and number of full baths, holding above grade living area constant, and between sale price and above grade living area, holding the number of full baths constant, is statistically significant at the 5% significance level, based upon the results of the hypothesis tests for the model coefficients.

The Q-Q plot of model #3 in Figure 20a indicates a departure of values from a normal distribution for lower and higher sale priced homes. The residual versus predictor plot of model #3 in Figure 20b demonstrates a heteroscedastic, non-constant variance, displayed as a widening column of plotted values as the number of full baths increase.

Figure 20: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #3. (b) Residual versus predictor plot of model #3.



(a) (b)

Log of Sale Price Response Models:

The models from the simple linear regression and multiple linear regression above were refit using the log transformation of sale price. The model summaries and parameters are provided and discussed. The goodness-of-fit of each model were assessed with two diagnostic plots, a Q-Q plot and a plot of the model residuals versus the predictor variable. The model training data set was used.

Model #4 Gr Liv Area:

Model #4 is a simple linear regression model of the log of sale price as a function of above grade living area.

Figure 21: Residuals for Model #4 Gr Liv Area

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-1.33124	-0.11884	0.01849	0.13568	0.74017

Figure 22: Coefficients for Model #4 Gr Liv Area

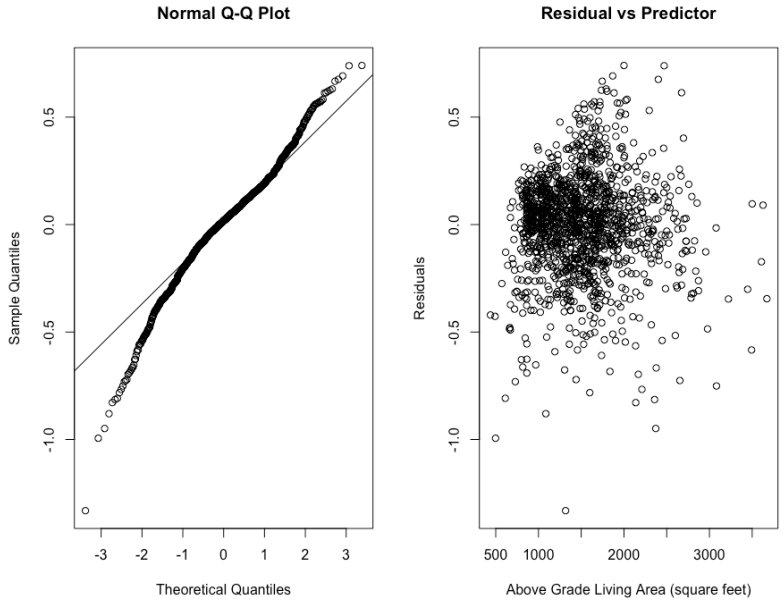
	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	11.17112378	0.2027143	551.08	< 2e-16
GrLivArea	0.00057460	0.00001286	44.67	< 2e-16

Figure 23: Model #4 summary

Residual standard error	0.2342 on 1402 degrees of freedom
Multiple R-squared	0.5873
Adjusted R-squared	0.587
F-statistic	1995 on 1 and 1402 DF
p-value	< 2.2e-16

From the model, the untransformed sale price is expected to be 0.05% higher ($\exp^{0.00057460}$) for each square foot increase in above grade living area. The residual standard error is 0.2342. The multiple R-squared value is 0.5873, which indicates that 58.5% of the variation in log transformed sale price (about its mean) can be explained by the variable above grade living area. The linear association between log transformed sale price and above grade living area is statistically significant at the 5% significance level, based upon the results of the hypothesis tests for the model coefficients.

Figure 24: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #4. (b) Residual versus predictor plot of model #4.



(a) (b)

The Q-Q plot of model #4 indicates a departure of values from a normal distribution more so for lower sale priced homes compared to higher sale priced homes. The residual versus predictor plot of model #4 demonstrates an improved homoscedastic variance as compared to Figure 12b.

Model #5 Full Bath:

Model #5 is a simple linear regression model of the log of sale price as a function of the number of full baths.

Figure 25: Residuals for Model #5 Full Bath

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-1.33493	-0.14610	0.00005	0.15025	1.53135

Figure 26: Coefficients for Model #5 Full Bath

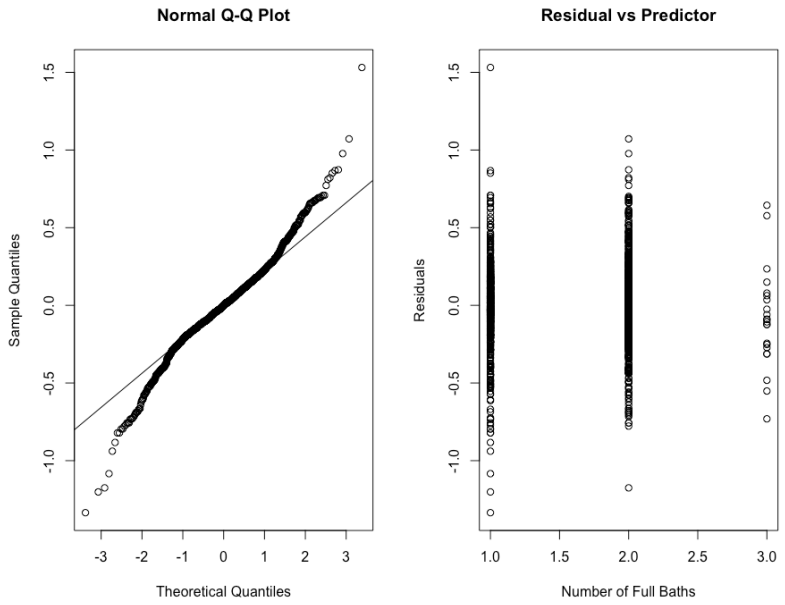
	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	11.34647	0.02227	509.45	< 2e-16
Full Bath	0.45156	0.01384	32.62	< 2e-16

Figure 27: Model #5 summary

Residual standard error	0.2749 on 1402 degrees of freedom
Multiple R-squared	0.4315
Adjusted R-squared	0.4311
F-statistic	1064 on 1 and 1402 DF
p-value	< 2.2e-16

From the model, the untransformed sale price is expected to be 57% higher ($\exp^{0.45156}$) for each full bath. The residual standard error is 0.2749. The multiple R-squared value is 0.4315, which indicates that 43.1% of the variation in sale price (about its mean) can be explained by the variable of number of full baths. The linear association between log transformed sale price and number of full baths is statistically significant at the 5% significance level, based upon the results of the hypothesis tests for the model coefficients.

Figure 28: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #5. (b) Residual versus predictor plot of model #5.



(a) (b)

The Q-Q plot of model #5 indicates an improvement of values for a normal distribution for higher sale priced homes compared to Figure 16a. The residual versus predictor plot of model #5 demonstrates a homoscedastic, constant variance as compared to Figure 16b.

The Q-Q plot of model #4 indicates a departure of values from a normal distribution more so for lower sale priced homes compared to higher sale priced homes. The residual versus predictor plot of model #4 demonstrates an improved homoscedastic, constant variance as compared to Figure 12b.

Multiple Linear Regression Model – Model #6:

Model #6 is a multiple linear regression model of the log of sale price as a function of the above grade living area and the number of full baths.

Figure 29: Residuals for Model #6

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-1.25577	-0.10376	0.01298	0.12665	0.95631

Figure 30: Coefficients for Model #6

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	11.25769457	0.02125335	529.690	< 2e-16
GrLivArea	0.00045157	0.00001641	27.522	< 2e-16
Factor(Full Bath)2	0.19601667	0.1529878	12.813	< 2e-16
Factor(Full Bath)3	0.14217456	0.05381989	2.642	0.00834

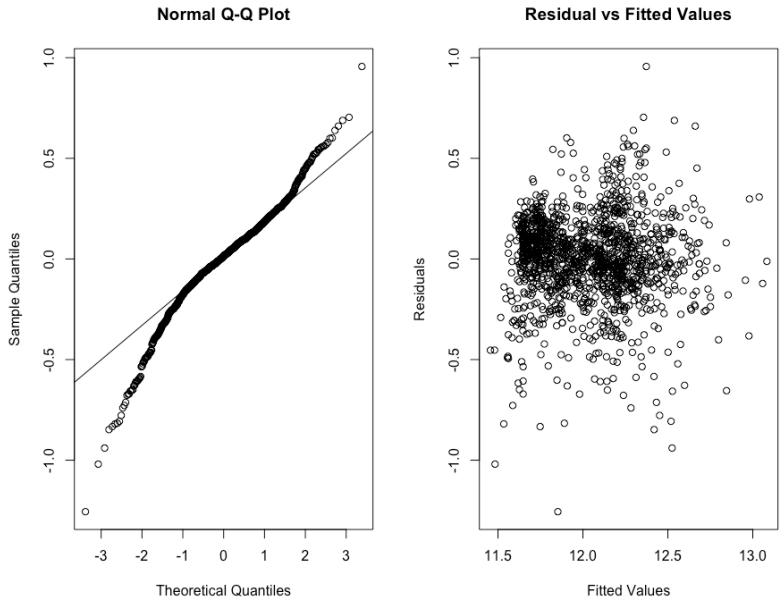
Figure 31: Model #6 summary

Residual standard error	0.2214 on 1400 degrees of freedom
Multiple R-squared	0.6318
Adjusted R-squared	0.6311
F-statistic	800.9 on 3 and 1400 DF
p-value	< 2.2e-16

From the model, the untransformed sale price is expected to be 0.04% higher ($\exp^{0.00045157}$) for each square foot increase in above grade living area when the number of full baths is held constant. The untransformed sale price is expected to be 21% higher ($\exp^{0.19601667}$) for a second full bath and to be 15% higher ($\exp^{0.19601667}$) for a third full bath when the above grade living area remains constant. The residual standard error is 0.2214. The multiple R-squared value is 0.6318, which indicates that 63.1% of the variation in log transformed sale price (about its mean) can be explained by a multiple linear regression association between the above grade living area and the number of full baths. The linear association between log transformed sale price and number of full baths, holding above grade living area constant, and between log transformed sale price and above grade living area, holding the number of full baths constant, is statistically significant at the 5% significance level, based upon the results of the hypothesis tests for the model coefficients.

The Q-Q plot of model #6 in Figure 32a indicates a departure of values from a normal distribution for lower and higher sale priced homes, although it is an improvement compared to the untransformed sale price model in Figure 20a. The residual versus predictor plot of model #6 in Figure 32b demonstrates a homoscedastic, constant variance, as compared to Figure 20b.

Figure 32: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #6. (b) Residual versus predictor plot of model #6.



(a) (b)

Conclusion:

Model #3 indicates that 60.8% of the variation in sale price (about its mean) can be explained by a multiple linear regression association between the above grade living area and the number of full baths. However, an assessment of the goodness-of-fit indicated that the residuals failed the zero mean assumption due to the heteroscedastic variance.

Model #6 indicates that 63.1% of the variation in log transformed sale price (about its mean) can be explained by a multiple linear regression association between the above grade living area and the of number of full baths. An improvement in the assessment of the goodness-of-fit was observed for the residuals in Figure 32b, although the normality regression assumption

References:

Benoit, K. (2011, March 17). Linear Regression Models with Logarithmic Transformations. Retrieved from http://kenbenoit.net/assets/courses/ME104/logmodels2.pdf

De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, Journal of Statistics Education, 19:3. Retrieved from http://amstat.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627#.WciC-WinFTE

De Cock, D. Ames Housing Data Documentation. Retrieved from https://ww2.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt

Pardoe, I. (2012). Applied Regression Modeling, Second Edition. Hoboken, NJ: John Wiley & Sons, Inc.

Stowell, S. (2014). Using R for Statistics. New York, NY: Apress.

Code:

```
# Jennifer Wanat
# Fall 2017
# Ames_assignment2.R

require(gridExtra)
#install.packages("sjPlot")
#require(sjPlot)
#require(Matrix)

path.name <- "~/Desktop/R/"
file.name <- paste(path.name, "ames_housing_data.csv", sep = "")

# Read in the csv file into an R data frame;
ames.df <- read.csv(file.name, header = TRUE, stringsAsFactors = FALSE)

# Show the header of the data frame;
head(ames.df)

# Show the structure of the data frame;
str(ames.df)

#Creating a waterfall of drop conditions
ames.df$dropCondition <- ifelse(ames.df$BldgType!='1Fam','01: Not SFR',
                          ifelse(ames.df$SaleCondition!='Normal','02: Non-Normal Sale',
                              ifelse(ames.df$Street!='Pave','03: Street Not Paved',
                                  ifelse(ames.df$GrLivArea >4000,'04: LT 4000 SqFt',
                                      ifelse(ames.df$LotArea >100000,'05: Lot 100000 SqFt',
                                          ifelse(ames.df$BedroomAbvGr <1, '06: No Bedrooms',
                                              ifelse(ames.df$FullBath <1, '07: No Full Baths',
                                          '99: Eligible Sample')
                                  ))))))

table(ames.df$dropCondition)

# Save the table
waterfall <- table(ames.df$dropCondition);

# Format the table as a column matrix for presentation;
as.matrix(waterfall,7,1)

# Eliminate all observations that are not part of the eligible sample population;
eligible.population <- subset(ames.df,dropCondition=='99: Eligible Sample');

# Check that all remaining observations are eligible;
```

```
table(eligible.population$dropCondition)

# Add a train/test flag to split the sample
eligible.population$u <- runif(n=dim(eligible.population)[1],min=0,max=1);
eligible.population$train <- ifelse(eligible.population$u<0.70,1,0);

# Check the counts on the train/test split
table(eligible.population$train)

# Check the train/test split as a percentage of whole
table(eligible.population$train)/dim(eligible.population)[1]

# Save the R data frame as an .RData object
saveRDS(eligible.population,file='/Users/jmwanat/Documents/Northwestern classes/MSPA
410/410 R/ames_assignment2.RData')

#EDA
table(eligible.population$GrLivArea!="NA")
table(eligible.population$GrLivArea!=0)
summary(eligible.population$GrLivArea)
quantile(eligible.population$GrLivArea)
#describe(eligible.population$GrLivArea)

addmargins(table(eligible.population$FullBath, useNA = c("always")))


par(mfrow = c(1,2))
plot(eligible.population$GrLivArea, eligible.population$SalePrice/1000,
    ylab = "Sale Price (000)",
    xlab = "Above Grade Living Area (square feet)")
abline(lm(eligible.population$SalePrice/1000 ~ eligible.population$GrLivArea), col =
"red", lwd = 2, lty = 2)
GrLivArea.loess <-
loess(eligible.population$SalePrice/1000~eligible.population$GrLivArea)
GrLivArea.predict <- predict(GrLivArea.loess)
lines(eligible.population$GrLivArea[order(eligible.population$GrLivArea)],
GrLivArea.predict[order(eligible.population$GrLivArea)],
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
hist(eligible.population$GrLivArea, xlab = "Above Grade Living Area (square feet)", col =
"lightgrey",
    ylim = c(0,800),
    main = "")
par(mfrow = c(1,1))

#boxplot of sale price versus full bath
```

```
boxplot(eligible.population$SalePrice/1000 ~ eligible.population$FullBath, col =
"lightgrey",
      ylab = "Sale Price (000)",
      xlab = "Number of Full Baths")
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")

#Change full bath variable from integer to factor
eligible.population$FullBath <- as.factor(eligible.population$FullBath)
#check data frame
str(eligible.population)

#descriptive statistics of sale price as a function of full bath variable
FullBath.min <- tapply(eligible.population$SalePrice, eligible.population$FullBath, min)
FullBath.median <- tapply(eligible.population$SalePrice, eligible.population$FullBath,
median)
FullBath.mean <- round(tapply(eligible.population$SalePrice,
eligible.population$FullBath, mean), 0)
FullBath.max <- tapply(eligible.population$SalePrice, eligible.population$FullBath, max)

overview <- cbind(FullBath.min, FullBath.median, FullBath.mean, FullBath.max)
colnames(overview) <- c("Min","Median", "Mean", "Max")
rownames(overview) <- c("1 Full Bath", "2 Full Bath", "3 Full Bath")
grid.table(overview)

#Quantile summary of sale price as a function of full bath
FullBath.quantile <- tapply(eligible.population$SalePrice, eligible.population$FullBath,
quantile)
FullBath.quantile



# Technically we should perform our EDA on the training data set
train.df <- subset(eligible.population,train==1);


# Fit a linear regression model with R
#Sale price as a function of Above Grade Living Area
model.1 <- lm(SalePrice ~ GrLivArea, data=train.df)

# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.1)
par(mfrow = c(1,1))

# Hint:  We need to check the assumptions of normality and homoscedasticity;
# (1) QQ Plot
```

```
# (2) Scatterplot of residuals versus predictor

par(mfrow = c(1,2))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.1$residuals)
qqline(model.1$residuals)
# Make a scatterplot
plot(train.df$GrLivArea,model.1$residuals,
    ylab = "Residuals",
    xlab = "Above Grade Living Area (square feet)")
title('Residual vs Predictor')
par(mfrow = c(1,1))

summary(model.1)
#sjt.lm(model.1)

# Fit a linear regression model with R
#Sale price as a function of Number of Full Baths
model.2 <- lm(SalePrice ~ as.numeric(FullBath), data=train.df)

# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.2)
par(mfrow = c(1,1))

par(mfrow = c(1,2))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.2$residuals)
qqline(model.2$residuals)
# Make a scatterplot
plot(as.numeric(train.df$FullBath),model.2$residuals,
    ylab = "Residuals",
    xlab = "Number of Full Baths")
title('Residual vs Predictor')
par(mfrow = c(1,1))

summary(model.2)

# Fit a linear regression model with R
#Sale price as a function of Above Grade Living Area and Number of Full Baths
model.3 <- lm(SalePrice ~ GrLivArea + factor(FullBath), data=train.df)

# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.3)
par(mfrow = c(1,1))
```

```
par(mfrow = c(1,2))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.3$residuals)
qqline(model.3$residuals)
# Make a scatterplot
plot(model.3$fitted.values,model.3$residuals,
    ylab = "Residuals",
    xlab = "Fitted Values")
title('Residual vs Fitted Values')
par(mfrow = c(1,1))

summary(model.3)

#Regression models for the tranformed response log(SalePrice)
#The models from above will be refit using log(SalePrice) as the response instead of
SalePrice
#log(Sale price) as a function of Above Grade Living Area
model.4 <- lm(log(SalePrice) ~ GrLivArea, data=train.df)

# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.4)
par(mfrow = c(1,1))

#We need to check the assumptions of normality and homoscedasticity;
# (1) QQ Plot
# (2) Scatterplot of residuals versus predictor

par(mfrow = c(1,2))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.4$residuals)
qqline(model.4$residuals)
# Make a scatterplot
plot(train.df$GrLivArea,model.4$residuals,
    ylab = "Residuals",
    xlab = "Above Grade Living Area (square feet)")
title('Residual vs Predictor')
par(mfrow = c(1,1))

summary(model.4)

# Fit a linear regression model with R
#log(Sale price) as a function of Number of Full Baths
model.5 <- lm(log(SalePrice) ~ as.numeric(FullBath), data=train.df)
```

```
# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.5)
par(mfrow = c(1,1))

par(mfrow = c(1,2))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.5$residuals)
qqline(model.5$residuals)
# Make a scatterplot
plot(as.numeric(train.df$FullBath),model.5$residuals,
    ylab = "Residuals",
    xlab = "Number of Full Baths")
title('Residual vs Predictor')
par(mfrow = c(1,1))

summary(model.5)

# Fit a linear regression model with R
#log(Sale price) as a function of Above Grade Living Area and Number of Full Baths
model.6 <- lm(log(SalePrice) ~ GrLivArea + factor(FullBath), data=train.df)

# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.6)
par(mfrow = c(1,1))

par(mfrow = c(1,2))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.6$residuals)
qqline(model.6$residuals)
# Make a scatterplot
plot(model.6$fitted.values,model.6$residuals,
    ylab = "Residuals",
    xlab = "Fitted Values")
title('Residual vs Fitted Values')
par(mfrow = c(1,1))

summary(model.6)
```