Assignment #1
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 needed to be conducted.

First, 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. Second, twenty variables from the data set were selected for a data quality check to look for errors and outliers. Descriptive statistics was performed on the variables when possible. Otherwise, the remaining variables were tabulated and examined for null entries or errors.

Then ten variables from the data quality check were selected for an initial EDA. Boxplots, histograms, and scatterplots were used to visualize the data relationships and distribution. The graphical representation can be collaborated against the data quality check. Finally, three variables from the EDA were selected for an initial EDA for modeling. The relationship of these three variables were investigated with the sale price and the log transformation of sale price. The EDA was 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

The resulting sample population data set (a data frame called eligible.population) contained 82 variables from 1992 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.

Data Quality Check:

Twenty variables were considered from the original 82 variables of the data set for a data quality check. If the variable was continuous, then a basic descriptive statistical and quantile summary of the variable was conducted. The descriptive statistical summary in Figure 3 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 4 includes the variable value at the 0, 25, 50, 75 and 100% quantiles.

If the variable was non-continuous (nominal, ordinal or discrete), then a table summarizing the counts of the factor levels was conducted. The exception was the discrete variable for the original construction date (YearBuilt), which had a descriptive statistical and quantile summary conducted, and this information was included with the continuous variables. All variables were checked for null values (NA), and continuous variables were checked for values equal to zero.

Figure 2: Twenty variables used for data quality check.

Variable	Description	Class
LotArea	Lot size in square feet	Continuous
Street	Type of road access to property	Nominal
Utilities	Type of utilities available	Ordinal
HouseStyle	Style of Dwelling	Nominal
YearBuilt	Original construction date	Discrete
RoofMat	Roof material	Nominal
Exterior1	Exterior covering on house	Nominal
BsmtFinType1	Rating of basement finished area	Ordinal
Heating	Type of heating	Nominal
CentralAir	Central air conditioning	Nominal
Electrical	Electrical system	Ordinal
GrLivArea	Above grade (ground) living area square feet	Continuous
FullBath	Full bathrooms above grade	Discrete
HalfBath	Half baths above grade	Discrete
BedroomAbvGr	Bedrooms above grade	Discrete
Fireplaces	Number of fireplaces	Discrete

GarageType	Garage location	Nominal
PavedDrive	Paved driveway	Ordinal
SaleCondition	Condition of sale	Nominal
SalePrice	Sale price in dollars	Continuous

Figure 3: Descriptive statistics.

Variable	Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
LotArea	2500	8125	9750	10580	11767	70761
YearBuilt	1872	1950	1968	1968	1996	2010
GrLivArea	334	1112	1445	1494	1759	3820
SalePrice	35000	130500	161875	178944	212075	625000

The statistics listed in Figure 3 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 4: Quantile summary.

Variable	0%	25%	50%	75%	100%
LotArea	2500	8125	9750	11767	70761
YearBuilt	1872	1950	1968	1996	2010
GrLivArea	334	1112	1445	1759	3820
SalePrice	35000	130500	161875	212075	625000

The output of the non-continuous variables are as follows. The column NA indicates if there were any observations that contained null values for the variable. The desired result for NA is zero. The exceptions were for the BsmtFin Type 1 variable, which had 45 NA counts, and for the GarageType variable, which had 73 NA counts. These NA entries could be due to an inadvertent missed entry, a typo, or the information was not specified.

Figure 5: Street data quality check.

Paved (Pave)	NA
1992	0

Figure 6: Utilities data quality check.

(AllPub)	(NoSewr)	NA
1991	1	0

Figure 7: House Style data quality check.

1.5 Fin	1.5 Unf	1Story	2.5Fin	2.5Unf	2Story	SFoyer	SLvl	NA
246	18	981	6	17	575	42	107	0

Figure 8: Roof Material data quality check.

CompShg	Membran	Metal	Tar&Grv	WdShake	WdShngl	NA
1964	1	1	13	7	6	0

Figure 9: Exterior covering on house (Exterior 1) data quality check.

AsbShng	25
BrkComm	5
BrkFace	62
CBlock	2
CemntBd	39
HdBoard	339
ImStucc	1
MetalSd	296
Plywood	128
PreCast	1
Stucco	34
VinylSd	681
Wd Sdng	333
WdShing	46
NA	0

Figure 10: Rating of basement finished area (BsmtFin Type 1) data quality check.

ALQ	BLQ	GLQ	LwQ	Rec	Unf	NA
317	222	516	116	225	551	45

Figure 11: Type of heating data quality check.

Floor	GasA	GasW	Grav	OthW	Wall	NA
1	1964	19	5	2	1	0

Figure 12: Central air conditioning data quality check.

No	Yes	NA
107	1885	0

Figure 13: Electrical system data quality check.

*	FuseA	FuseF	FuseP	SBrkr	Mix*	NA
1	142	29	4	1816	*	0

*There was no description for the entry with 1 count, and the Mix category was not specified

Figure 14: Full bath data quality check.

0	1	2	3	NA
5	997	954	36	0

Figure 15: Half bath data quality check.

0	1	2	NA
1240	747	5	0

Figure 16: Bedrooms above grade data quality check.

0	1	2	3	4	5	NA
4	38	412	1232	278	28	0

Figure 17: Fireplaces data quality check.

0	1	2	3	4	NA
913	909	160	9	1	0

Figure 18: Garage type data quality check.

2Types	Attchd	Basment	BuiltIn	CarPort	Detchd	NA
11	1203	19	124	5	557	73

Figure 19: Paved drive data quality check.

N (Dirt/Gravel)	P (Partial Pavement)	Y (Paved)	NA
147	56	1789	0

Figure 20: Sale condition data quality check.

Normal	NA
1992	0

Initial Exploratory Data Analysis:

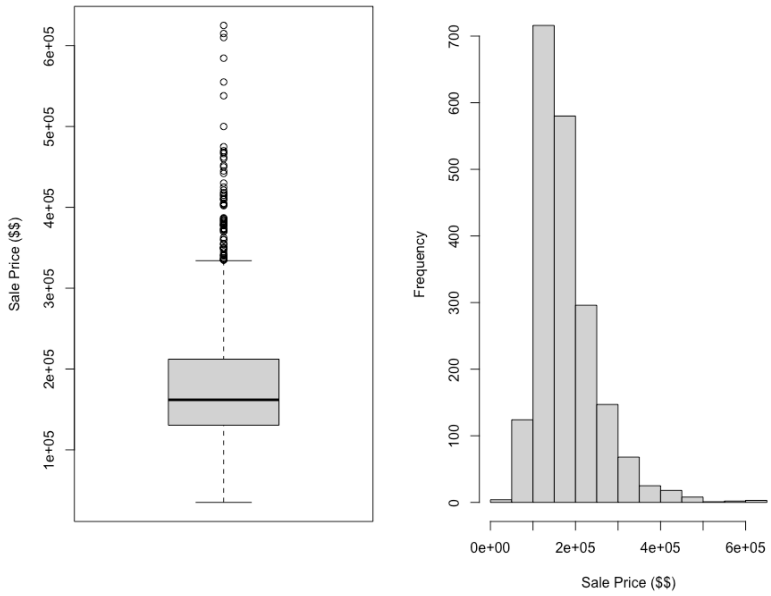
Ten variables were selected for an initial exploratory data analysis from the twenty variables used in the data quality check. The ten variables are listed in Figure 21.

Figure 21: Ten variables used for initial exploratory data analysis.

Variable	Description	Class
LotArea	Lot size in square feet	Continuous
Utilities	Type of utilities available	Ordinal
YearBuilt	Original construction date	Discrete
GrLivArea	Above grade (ground) living area square feet	Continuous
FullBath	Full bathrooms above grade	Discrete
HalfBath	Half baths above grade	Discrete
BedroomAbvGr	Bedrooms above grade	Discrete
GarageType	Garage location	Nominal
Heating	Type of heating	Nominal
SalePrice	Sale price in dollars	Continuous

The initial exploratory data analysis utilized boxplots, histograms and scatterplots, as appropriate.

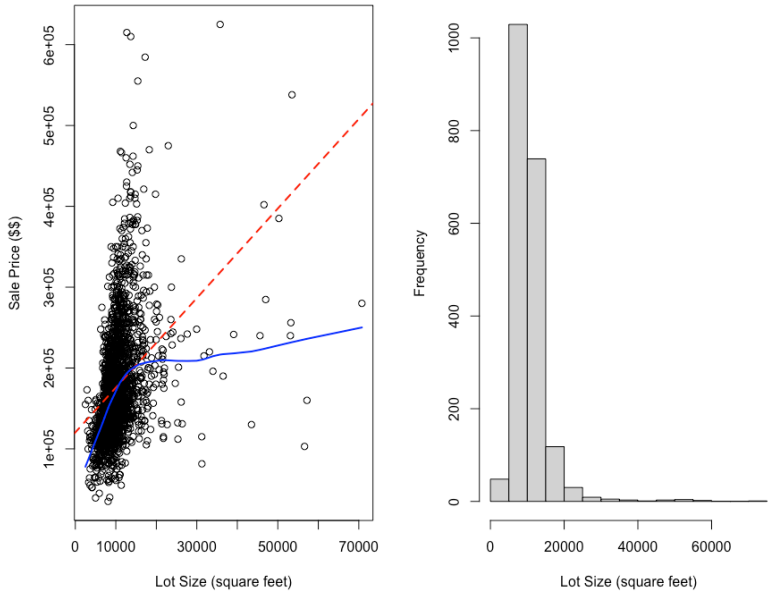
Figure 22: (a) Boxplot and (b) histogram of sale price.



(a) (b)

The majority of sale prices are less than $200,000. The house sale prices are positively skewed, which explains why the mean is greater than the median (Figure 3).

Figure 23: (a) Scatterplot of sale price versus lot size. The dashed line was estimated using ordinary least squares or OLS. The solid line is a lowess smoother. (b) Histogram of lot size.

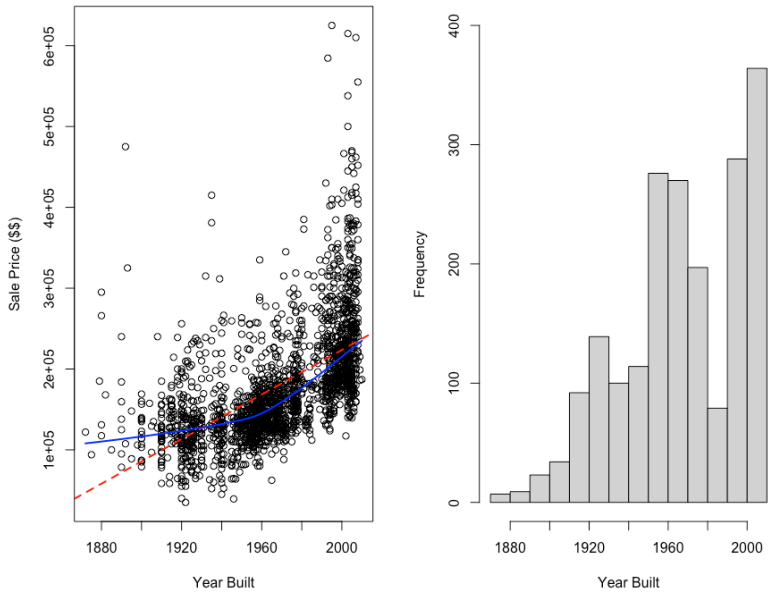


(a) (b)

The lowess smoother line and the regression line have a poor agreement between the two lines in the scatterplot. Both lines have a positive slope until a lot size of approximately 15,000 square feet. The lowess smoother line for lot sizes larger than that levels off. The plotted values of sale price and lot size display a heteroscedastic, non-constant variance, for small and large lot sizes. Visually, there is little correlation between sale price and lot size.

An examination of the histogram indicates that the majority of lot sizes are less than 20,000 square feet. The lot sizes are positively skewed, which is also supported by the data in Figure 3 in which the mean is greater than the median (Figure 3).

Figure 24: (a) Scatterplot of sale price versus year built. The dashed line was estimated using ordinary least squares or OLS. The solid line is a lowess smoother. (b) Histogram of year built.

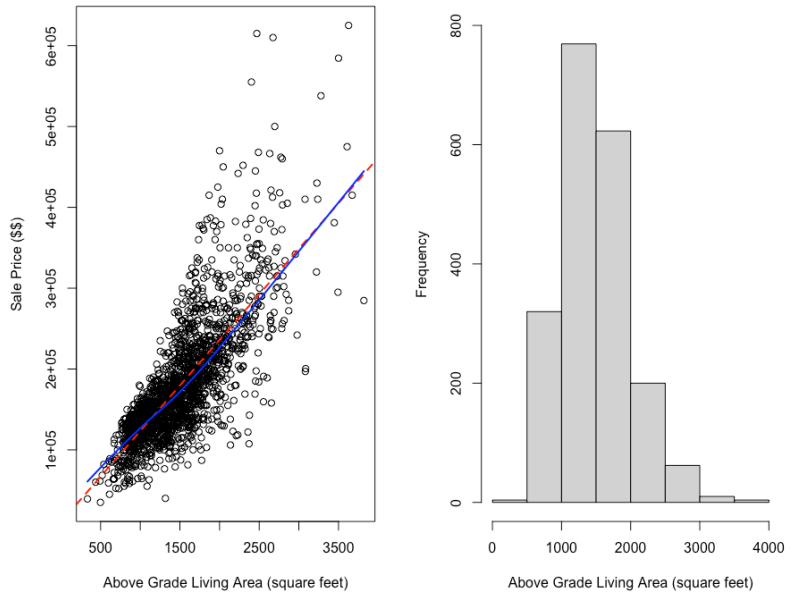


(a) (b)

The lowess smoother line and the regression line both have an upward slope for values greater than the year 1940 in the scatterplot. Both lines agree less for older homes. The plotted values of sale price and year built display a heteroscedastic, non-constant variance, for older and newer homes. Home sale prices have been generally increasing since the 1960s.

An examination of the histogram indicates that the number of homes built by decade has generally increased with a notable exception between the years of 1980-1990, when the Midwest was affected by a farming recession. The highest number of homes were built between the years of 2000-2010.

Figure 25: (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 lowess smoother. (b) Histogram of above grade living area.



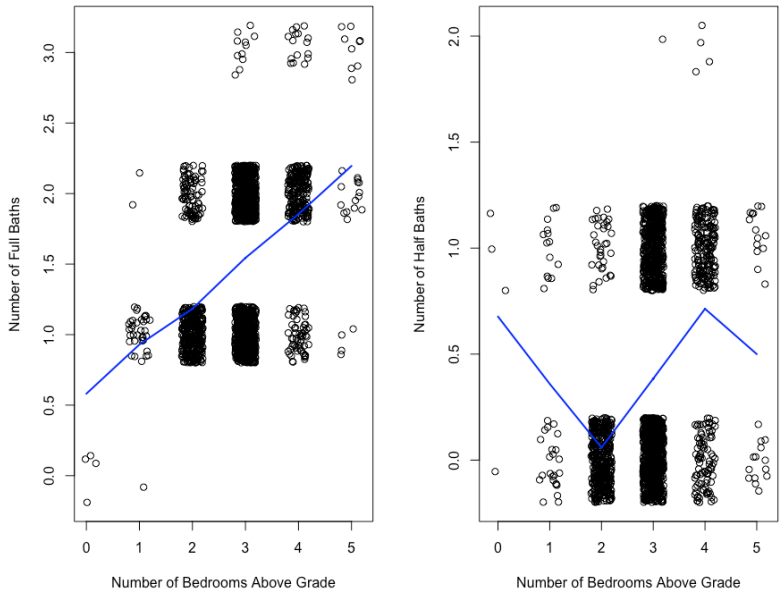
(a) (b)

The lowess 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. A transformation of both variables may be useful.

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

Figures 26 and 27 indicates that the number of bathrooms increase with bedrooms. Most homes have one full bathroom (740 total). Most homes do not have a half-bath (1240 total).

Figure 26: (a) Scatterplot of number of full baths versus number of bedrooms above grade. The solid line is a lowess smoother. (b). Scatterplot of number of half baths versus number of bedrooms above grade. The solid line is a lowess smoother.



(a) (b)

Figure 27: Cross tabulation of number of full baths (vertical) versus number of half baths (horizontal).

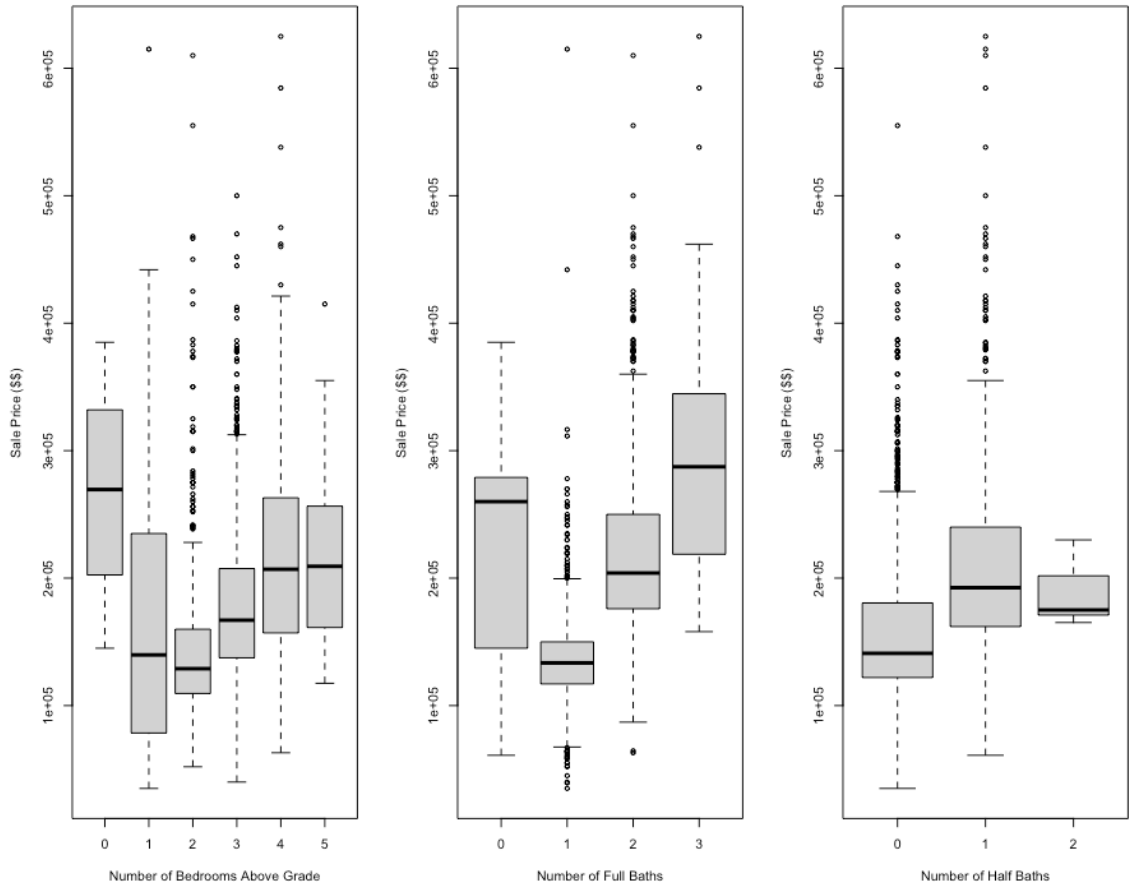
	0	1	2	Sum
0	1	4	0	5
1	740	252	5	997
2	480	474	0	954
3	19	17	0	36
Sum	1240	747	5	1992

Figure 16 indicates that there are four homes with zero bedrooms above grade. Generally, the sale price increases as the number of bedrooms increase, as can be observed in Figure 28a. The median sale price is nearly identical for homes with four or five bedrooms.

Figure 27 indicates that there are five homes with zero full baths. Generally, the sale price increases as the number of full baths increases, as can be observed in Figure 28b.

Figure 27 indicates that there are 1240 homes with zero half baths. Figure 28c demonstrates that there is a wide variability in sale price as the number of half baths increases.

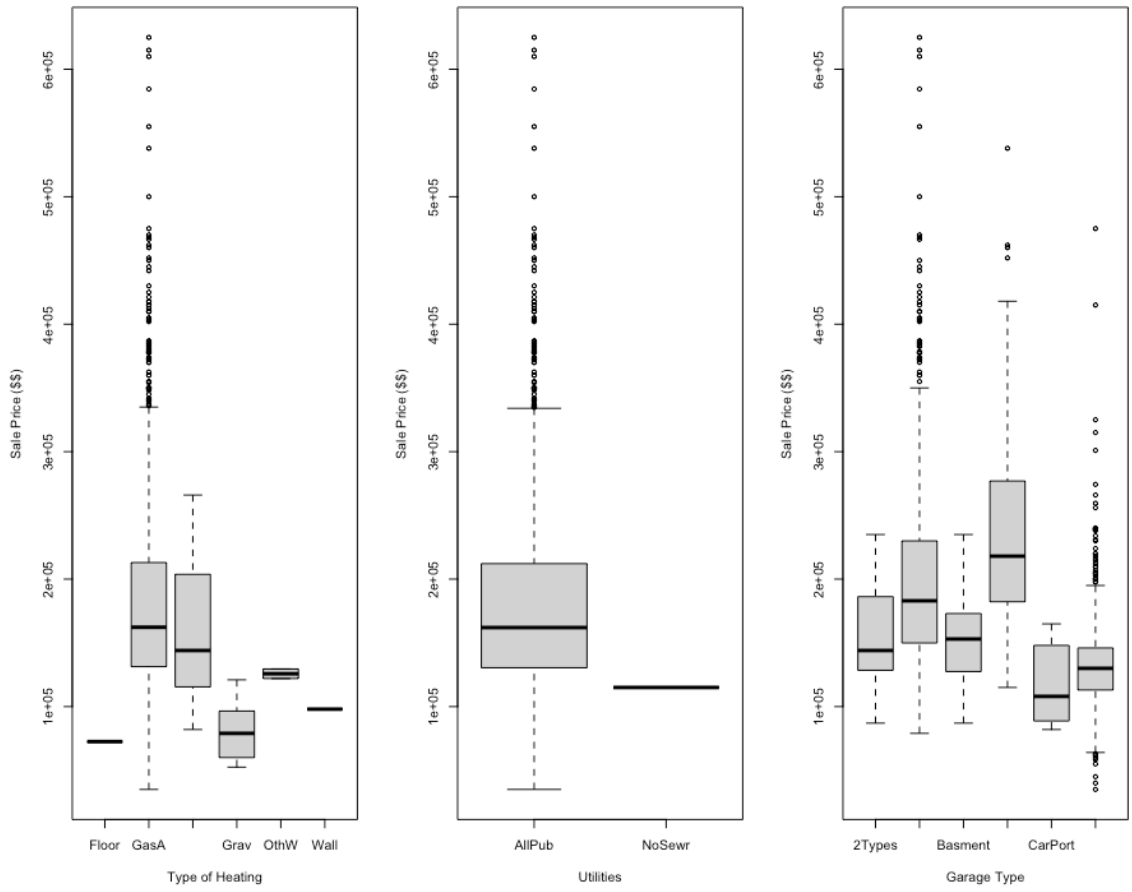
Figure 28: (a) Boxplot of sale price versus number of bedrooms above grade. (b) Boxplot of sale price versus number of full baths. (c) Boxplot of sale price versus number of half baths.



(a) (b) (c)

In figure 29, there is a wide variation in sale price for homes with GasA (gas forced warm air furnace) heating. There is also a wide variation in sale price for homes with AllPub (all public utilities E, G, W, & S). It is also noted that there is a wide variation in sale price for homes with attached, built-in, and detached from home garages.

Figure 29: (a) Boxplot of sale price versus type of heating. (b) Boxplot of sale price versus utilities. (c) Boxplot of sale price versus garage type.



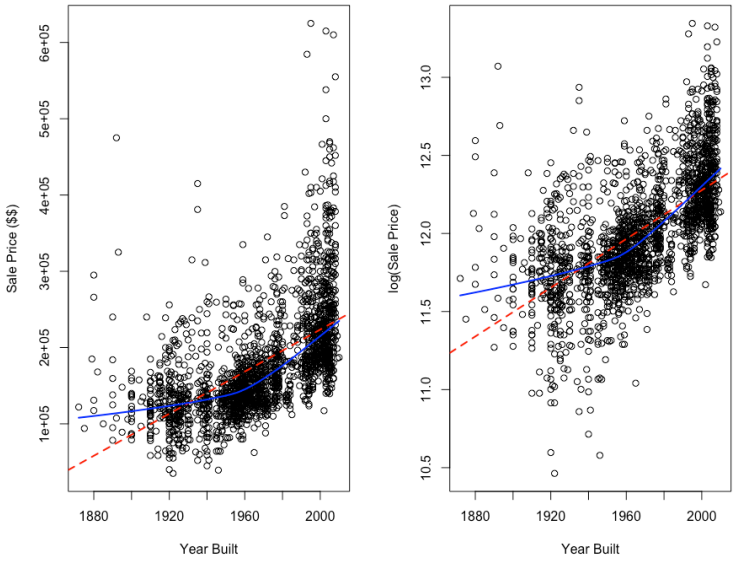
(a) (b) (c)

Exploratory Data Analysis for Modeling:

Three variables from the ten variables in the initial exploratory data analysis were chosen to explore their relationship between sale price and the log transformation of sale price. The three variables chosen were year built, above grade living area, and number of bedrooms. The values of sale price and log(sale price) versus each variable were graphed in a scatterplot, and a regression line and lowess smoothing line were added.

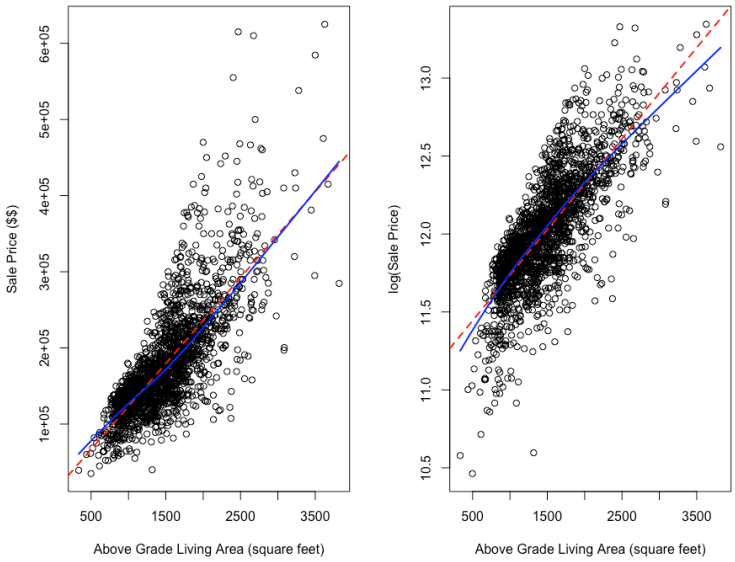
In Figure 30, the log of sale price lowered the spread of values for newer homes, but caused a greater variation of values for homes built between the years of 1920 and 1960. A log transformation of sale price does not change the spatial relation between the regression and lowess smoother lines.

Figure 30: (a) Scatterplot of sale price versus year built. The dashed line was estimated using ordinary least squares or OLS. The solid line is a lowess smoother. (b) Scatterplot of log transformation of sale price versus year built. The dashed line was estimated using ordinary least squares or OLS. The solid line is a lowess smoother.



(a) (b)

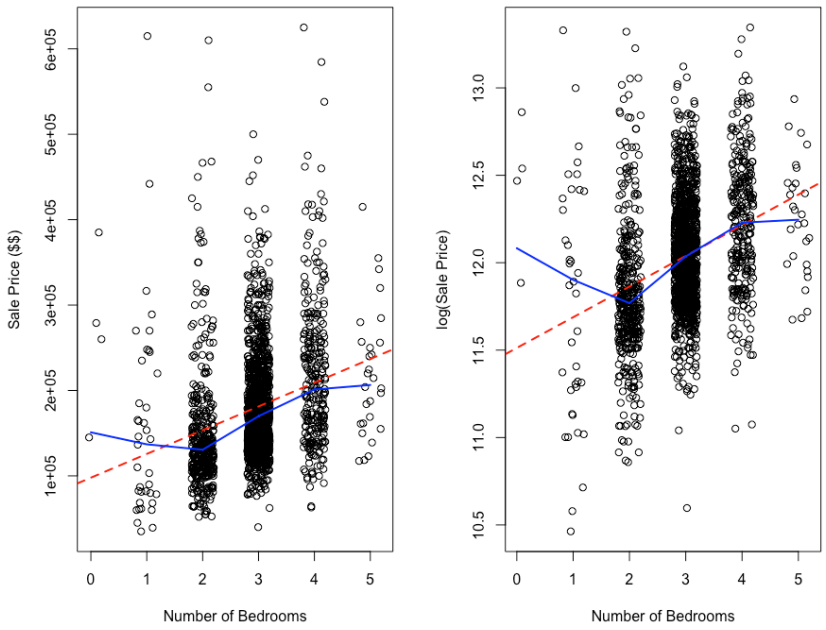
Figure 31: (a) Scatterplot of sale price versus above grade living area. The dashed line was estimated using ordinary least squares or OLS. The solid line is a lowess smoother. (b) Scatterplot of log transformation of sale price versus above grade living area. The dashed line was estimated using ordinary least squares or OLS. The solid line is a lowess smoother.



(a) (b)

Additionally, a log transformation of sale price did not change the spatial relation between the regression and lowess smoother lines for either above grade living area or the number of bedrooms. A log transformation of sale price does not appear to have any added benefit with the three variables examined.

Figure 32: (a) Scatterplot of sale price versus number of bedrooms. The dashed line was estimated using ordinary least squares or OLS. The solid line is a lowess smoother. (b) Scatterplot of log transformation of sale price versus number of bedrooms. The dashed line was estimated using ordinary least squares or OLS. The solid line is a lowess smoother.



(a) (b)

Conclusion:

There appears to be a strong correlation between sale price and above grade living area. The number of bedrooms may also be a useful predictor, especially if homes with zero bedrooms are dropped from the data set. Another variable that may be useful could be the number of full baths. Finally, the build year after 1960 has a positive correlation with sale price.

References:

Dean De Cock. (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

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

Code:

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

install.packages("Hmisc")
require(Hmisc)
require(gridExtra)

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)


# This plot was created to view the 5 outliers and unusual values mentioned
# in the data documentation special notes section
plot(ames.df$GrLivArea, ames.df$SalePrice,
    xlab = "Above Grade Living Area (square feet)", ylab = "Sale Price")


#Creating a waterfall of drop conditions
# Single ifelse() statement
# ifelse(condition, value if condition is TRUE, value if the condition is FALSE)

# Nested ifelse() statement
# ifelse(condition1, value if condition1 is TRUE,
#         ifelse(condition2, value if condition2 is TRUE,
#         value if neither condition1 nor condition2 is TRUE
#         )
# )


# Create a waterfall of drop conditions;
# Work the data frame as a 'table' like you would in SAS or SQL;
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',
                            '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,6,1)

data[,"waterfall", drop=FALSE]

# 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)


#Pick twenty variables and run a data quality check on them;

table(eligible.population$LotArea!="NA")
table(eligible.population$LotArea!=0)
summary(eligible.population$LotArea)
quantile(eligible.population$LotArea)

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

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

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

table(eligible.population$YearBuilt!="NA")
table(eligible.population$YearBuilt, useNA = c("always"))
summary(eligible.population$YearBuilt)
quantile(eligible.population$YearBuilt)

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

table(eligible.population$Exterior1, useNA = c("always"))

table(eligible.population$BsmtFinType1, useNA = c("always"))
```

```
table(eligible.population$Heating, useNA = c("always"))

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

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

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

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

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

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

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

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

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

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

table(eligible.population$SalePrice!="NA")
summary(eligible.population$SalePrice)
quantile(eligible.population$SalePrice)
describe(eligible.population$SalePrice)


#Pick ten variables from the twenty variables from the data
#quality check to explore in initial exploratory data analysis;

#The ten variables are:
#lot area, utilities, year built, GrLivArea, Full bath
#half bath, bedroom, garage type, heating, sale price

#Continuous: lot area, year built, GrLivArea, sale price
#Discrete: utilities, full bath, half bath, bedroom, garage type, heating

par(mfrow = c(1,2))
boxplot(eligible.population$SalePrice, ylab = "Sale Price ($$)", col = c("lightgrey"),
      coef = 3.0, do.conf = TRUE, do.out = TRUE)
hist(eligible.population$SalePrice, xlab = "Sale Price ($$)", col = "lightgrey",
```

```
    main = "")
par(mfrow = c(1,1))

par(mfrow = c(1,2))
plot(eligible.population$LotArea, eligible.population$SalePrice,
    ylab = "Sale Price ($$)",
    xlab = "Lot Size (square feet)")
abline(lm(eligible.population$SalePrice ~ eligible.population$LotArea), col = "red", lwd =
2, lty = 2)
lines(lowess(eligible.population$LotArea, eligible.population$SalePrice),
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010", outer = TRUE)
hist(eligible.population$LotArea, xlab = "Lot Size (square feet)", col = "lightgrey",
    main = "")
par(mfrow = c(1,1))

par(mfrow = c(1,2))
plot(eligible.population$YearBuilt, eligible.population$SalePrice,
    ylab = "Sale Price ($$)",
    xlab = "Year Built")
abline(lm(eligible.population$SalePrice ~ eligible.population$YearBuilt), col = "red", lwd
= 2, lty = 2)
lines(lowess(eligible.population$YearBuilt, eligible.population$SalePrice),
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010", outer = TRUE)
hist(eligible.population$YearBuilt, xlab = "Year Built", col = "lightgrey",
    ylim = c(0,400),
    main = "")
par(mfrow = c(1,1))

par(mfrow = c(1,2))
plot(eligible.population$GrLivArea, eligible.population$SalePrice,
    ylab = "Sale Price ($$)",
    xlab = "Above Grade Living Area (square feet)")
abline(lm(eligible.population$SalePrice ~ eligible.population$GrLivArea), col = "red", lwd
= 2, lty = 2)
lines(lowess(eligible.population$GrLivArea, eligible.population$SalePrice),
     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))

par(mfrow = c(1,2))
```

```
table(eligible.population$BedroomAbvGr, eligible.population$FullBath)
plot(jitter(eligible.population$BedroomAbvGr), jitter(eligible.population$FullBath),
    ylab = "Number of Full Baths",
    xlab = "Number of Bedrooms Above Grade")
lines(lowess(eligible.population$BedroomAbvGr, eligible.population$FullBath),
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
table(eligible.population$BedroomAbvGr, eligible.population$HalfBath)
plot(jitter(eligible.population$BedroomAbvGr), jitter(eligible.population$HalfBath),
    ylab = "Number of Half Baths",
    xlab = "Number of Bedrooms Above Grade")
lines(lowess(eligible.population$BedroomAbvGr, eligible.population$HalfBath),
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
par(mfrow = c(1,1))

par(mfrow = c(1,3))
boxplot(eligible.population$SalePrice ~ eligible.population$BedroomAbvGr, col =
"lightgrey",
       ylab = "Sale Price ($$)",
       xlab = "Number of Bedrooms Above Grade")
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
boxplot(eligible.population$SalePrice ~ eligible.population$FullBath, col = "lightgrey",
       ylab = "Sale Price ($$)",
       xlab = "Number of Full Baths")
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
boxplot(eligible.population$SalePrice ~ eligible.population$HalfBath, col = "lightgrey",
       ylab = "Sale Price ($$)",
       xlab = "Number of Half Baths")
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
par(mfrow = c(1,1))



par(mfrow = c(1,3))
boxplot(eligible.population$SalePrice ~ eligible.population$Heating, col = "lightgrey",
       ylab = "Sale Price ($$)",
       xlab = "Type of Heating")
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
boxplot(eligible.population$SalePrice ~ eligible.population$Utilities, col = "lightgrey",
       ylab = "Sale Price ($$)",
       xlab = "Utilities")
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
boxplot(eligible.population$SalePrice ~ eligible.population$GarageType, col = "lightgrey",
       ylab = "Sale Price ($$)",
       xlab = "Garage Type")
```

```
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
par(mfrow = c(1,1))

grid.table(addmargins(table(eligible.population$FullBath, eligible.population$HalfBath,
dnn = c("FullBath", "HalfBath"))))

#Pick three variables from the ten variables from the initial exploratory data analysis
#Explore their relationship with SalePrice and Log(SalePrice)

par(mfrow = c(1,2))
plot(eligible.population$YearBuilt, eligible.population$SalePrice,
    ylab = "Sale Price ($$)",
    xlab = "Year Built")
abline(lm(eligible.population$SalePrice ~ eligible.population$YearBuilt), col = "red", lwd
= 2, lty = 2)
lines(lowess(eligible.population$YearBuilt, eligible.population$SalePrice),
     col = "blue", lwd = 2, lty = 1)
#log of sale price
plot(eligible.population$YearBuilt, log(eligible.population$SalePrice),
    ylab = "log(Sale Price)",
    xlab = "Year Built")
abline(lm(log(eligible.population$SalePrice) ~ eligible.population$YearBuilt), col = "red",
lwd = 2, lty = 2)
lines(lowess(eligible.population$YearBuilt, log(eligible.population$SalePrice)),
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010", outer=TRUE)
par(mfrow = c(1,1))

par(mfrow = c(1,2))
plot(eligible.population$GrLivArea, eligible.population$SalePrice,
    ylab = "Sale Price ($$)",
    xlab = "Above Grade Living Area (square feet)")
abline(lm(eligible.population$SalePrice ~ eligible.population$GrLivArea), col = "red", lwd
= 2, lty = 2)
lines(lowess(eligible.population$GrLivArea, eligible.population$SalePrice),
     col = "blue", lwd = 2, lty = 1)
#log of GrLivArea
plot(eligible.population$GrLivArea, log(eligible.population$SalePrice),
    ylab = "log(Sale Price)",
    xlab = "Above Grade Living Area (square feet)")
abline(lm(log(eligible.population$SalePrice) ~ eligible.population$GrLivArea), col = "red",
lwd = 2, lty = 2)
lines(lowess(eligible.population$GrLivArea, log(eligible.population$SalePrice)),
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010", outer=TRUE)
par(mfrow = c(1,1))
```

```
par(mfrow = c(1,2))
plot(jitter(eligible.population$BedroomAbvGr), eligible.population$SalePrice,
    ylab = "Sale Price ($$)",
    xlab = "Number of Bedrooms")
abline(lm(eligible.population$SalePrice ~ eligible.population$BedroomAbvGr), col =
"red", lwd = 2, lty = 2)
lines(lowess(eligible.population$BedroomAbvGr, eligible.population$SalePrice),
     col = "blue", lwd = 2, lty = 1)
#log with bedroom
plot(jitter(eligible.population$BedroomAbvGr), log(eligible.population$SalePrice),
    ylab = "log(Sale Price)",
    xlab = "Number of Bedrooms")
abline(lm(log(eligible.population$SalePrice) ~ eligible.population$BedroomAbvGr), col =
"red", lwd = 2, lty = 2)
lines(lowess(eligible.population$BedroomAbvGr, log(eligible.population$SalePrice)),
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010", outer = TRUE)
par(mfrow = c(1,1))
```