Housing Data Analysis

AMS 572 Group 2 Report

Group Members: Bohong Huang
Jing Hao
Joanna Wong
Nick Handelman
Yiming Wan

Introduction

Finding a house is a common task that everyone will have to face. To find a home, thorough research has to be done that will fit the needs of the buyer or for the seller to the worth of their house. There are many house characteristics that determine a house such as the age of the house and the number of bedrooms and bathrooms. The data can also be seen and tested for relationships for different characters in different perspectives: the buyer, seller and developer. Since the group members are from different backgrounds, there needed to be data that could be easily understood by everyone. The group found a common feature, which was everyone had a home.

Description of Data

The data set chosen was "Housing Prices" from the "Data and Story Library"[1]. The random sample involves 1057 samples of houses in Saratoga, New York. For each sample, the price, living area, number of bathrooms, number of bedrooms, number of fireplaces, lot size and age were collected. The initial purpose of the data set was to determine the home values based on all the information collected. For the project, the data was modified by removing the 8th column of the binary data involving whether the house contained a fireplace or not. Also, the initial investigation of the data revealed that 13 samples contained value of "NA" or “0” for the lot size, so those data samples were removed. Among the remaining 1044 samples, there is no obviously erroneous data.

Questions of Interest

1.Analysis with Independent Samples T-Test

Every buyer wants to spend less money to buy a better house, and the prices of houses vary with many factors. The data indicates that the house price may be related to the age of house. Therefore, the first question of interest is: "Is the mean price of the newest half of the houses the same as the mean price of the oldest half of the houses?" The median age of the houses in the sample is 18 years, so the data were split into two groups. Group 1 consists of the houses newer than or equal to 18 years (564 samples). Group 2 consists of houses older than 18 years (480 samples). R code is provided in Appendix 2 and R output is provided in Appendix 3.

1.1.Hypotheses

$Ho : \mu_1 = \mu_2$

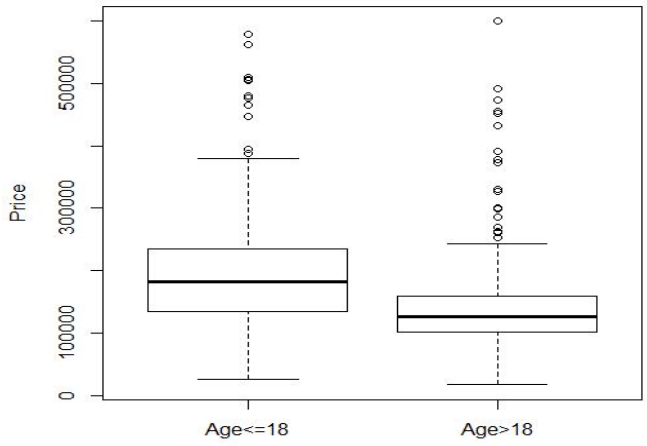
$Ha : \mu_1 \neq \mu_2$

The significant level we choose is $\alpha = 0.05$

[1] http://lib.stat.cmu.edu/DASL/.

1.2 Visual Analysis

Boxplots of Price vs Age <=18 and Price vs. Age >18



The box plots indicate that the median house less than or equal to 18 years is more expensive than the median house older than 18 years. For both groups, the data appear to be highly skewed and consists of many outliers, so it is difficult to make a statement about the mean house in each group. Also, it is not clear that the group variances are equal.

1.3 Validity of Assumptions

The data is a random sample, so the observations are independent of one another. Shapiro-Wilk tests on each of the groups (yellow highlight) indicate that both populations are non-normal and there are many outliers. However, due to the large sample sizes and the Central Limit Theorem, the normality assumption for the distribution of means for both groups is still valid. An equal variances F-test on the two groups (orange highlight) indicates that the variances are not equal. However, the validity of this test is questionable since the populations are non-normal. Considering the visual analysis and lacking any other evidence to the contrary, the t-test will be run assuming unequal variances.

1.4 Conclusion

Since we assumed unequal variances, a Welch Independent Samples T- test was performed, and the results are highlighted in green. With a p-value is less than 0.00000000000000022, H_0 is rejected at $\alpha = 0.05$. The 95% confidence interval is (46034.43, 63352.19), which provides evidence that the mean price for houses that are less than or equal to 18 years is greater than the mean price for the houses that are older than 18 years.

2. Analysis with One-Way ANOVA

There are many preferences that a home buyer has such as if they want a fireplace or not. A possible criteria a buyer could have would be a strong preference for the number of fireplaces in the house and a strong preference for the age of the house. Considering this, a question of interest is "For houses with 0, 1, 2, ..., n fireplaces, is the mean age equal?" Analysis of the data indicates that n = 4 with the following counts:

Fireplaces:	0	1	2	3	4
House Count:	422	595	24	2	1

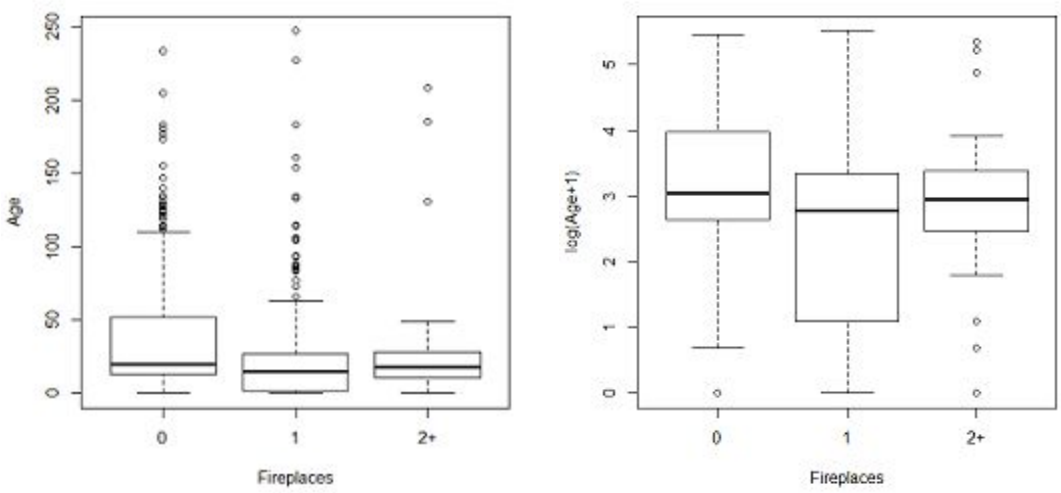
A one-way ANOVA is an appropriate test to address the question, assuming the data meet the required assumptions. Considering the group counts, the ANOVA test will have 3 groups: houses with 0 fireplaces, houses with 1 fireplace and houses with 2 or more fireplaces. It turns out that the raw data do not meet the required assumptions, so a one-way ANOVA is run on the log(x+1) transformed data. R code is provided in Appendix 4 and R output is provided in Appendix 5.

2.1 Hypotheses

Raw Data	log(x+1) Transformed Data
Ho : $\mu_0 = \mu_1 = \mu_{2+}$ Ha : at least one μ does not equal the others	Ho : $\mu_0' = \mu_1' = \mu_{2+}'$ Ha : at least one μ' does not equal the others

The significant level we choose for both tests is $\alpha = 0.05$.

2.2 Visual Analysis

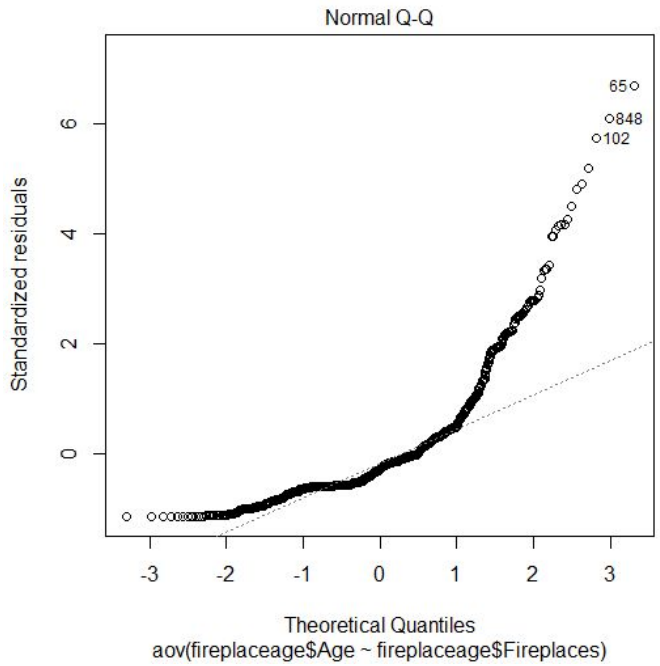


For the raw data (boxplot on the left), it is difficult to make an initial statement about the equality of the group population means based on the data. Within each group, the samples are clearly not normally distributed, indicating that their respective populations may not be normally distributed. The data does indicate that, due to the right skewness, the population means are likely greater than the population medians.

The transformed samples indicate that the transformed populations are more normally distributed within each group. This indicates that the transformed means are closer to the transformed medians. Considering that the sample medians are close, it is easier to make an initial statement that the transformed group means are equal.

2.3 One-Way ANOVA on the Raw Data

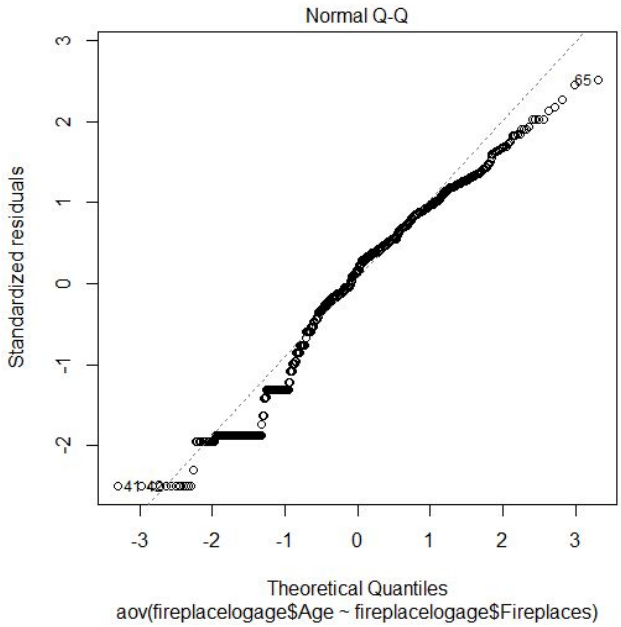
The test assumes that the group variances are equal. The boxplots indicate that the groups are not all normally distributed, and this is confirmed by a Shapiro-Wilk test on each group (yellow highlight). So, Levene's Test is used to assess the homogeneity of variances across the groups. The results (green highlight) indicate that the groups do not have the same variance.



After running the ANOVA, the assumption of normally distributed residuals can be addressed. The normal plot given above shows that residuals are not normally distributed. This is confirmed by a Shapiro-Wilk test on the residuals (light blue highlight). With an F-statistic of 37.23 on 2 and 1041 degrees of freedom, H_0 is rejected. However, since the required assumptions are not met, the result is questionable. It is possible that a transformation of the data can be applied, and the resulting transformed data will meet the assumptions.

2.4 One-Way ANOVA on the log(x+1) Transformed Data

Due to the great difference in house ages (minimum 0 years and maximum 247 years) a log transformation is appropriate. However, several samples have an age of 0, so the log transformation is taken on the age+1. Considering the boxplots, it does appear that the transformation improved the normality of the data. The result of a Shapiro-Wilk test (blue highlight) on group 2+ does not provide enough evidence to reject that its population is normally distributed. However, Shapiro-Wilk tests (blue highlight) on groups 0 and 1 do reject that those group populations are normally distributed. Since the groups are not all normally distributed, Levene's Test is used to assess the homogeneity of variances across the groups. The result (magenta highlight) indicate that the groups do not have the same variance.



After running the test, the assumption of normally distributed residuals can be addressed. The normal plot given above shows that residuals are not normally distributed, but does appear to be more normally distributed than the residuals from the raw data. A Shapiro-Wilk test on the residuals (orange highlight) confirms the non-normality. With an F-statistic of 50.09 on 2 and 1041 degrees of freedom, *Ho* is rejected. That is, at least one of the means of the log transformed data are not equal to the others. Since the required assumptions are not met, this result is questionable, but may be more valid than the results of the test on the raw data.

3. Analysis with Multiple Linear Regression

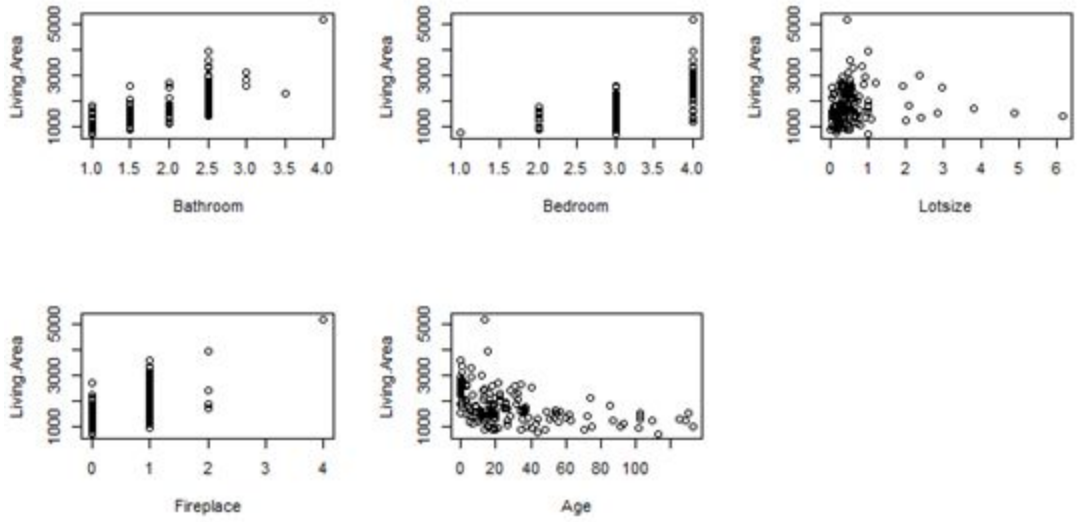
For house constructors, determination of the living area is a consideration. Looking at the data available, it appears that bathrooms(1), bedrooms(2), fireplaces(3), lot size(4) and age(5) could all have an influence on the total living area. A multiple linear regression model is appropriate to address this concern, assuming the data meet the required assumptions. Initially, all 5 variables are considered, and the results indicate that age is not a significant contributing factor and that a transformation may help meet certain assumptions. For the second model, the other 4 variables are considered and a log transform is applied to the dependent variable (living area) in an attempt to develop a model that meets the required assumptions.

3.1 Hypotheses

Raw Data with 5 variables	Log Transformed DV with 4 Raw Variables
Ho : $\beta_i = 0$ for all i Ha : at least one $\beta_i \neq 0$	Ho : $\beta''_i = 0$ for all i Ha : at least one $\beta''_i \neq 0$

The significant level we choose for all tests is $\alpha = 0.05$.

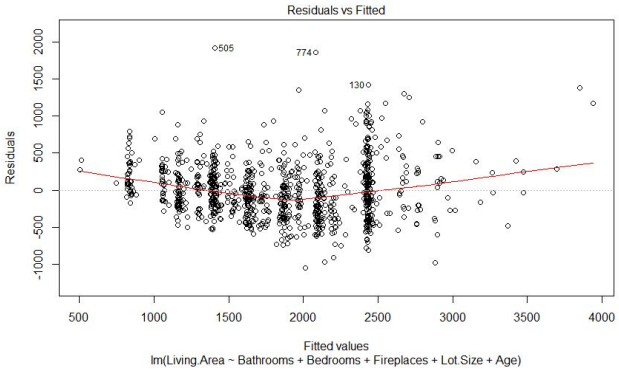
3.2 Visual Analysis of Simple Relationships

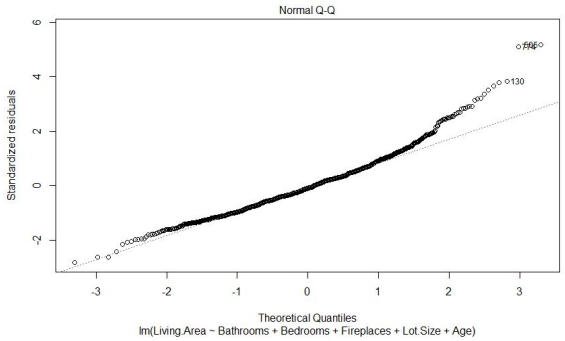


From the above plots, it appears that living area has a positive linear relationship with bathrooms and fireplaces and possibly bedrooms. Living area appears to have a slight negative curvilinear relationship with age, but this may just be due to the visual influence of certain outliers. For lot size, most observations appear to be in one cluster, and the remaining observations give a slight curvilinear appearance.

3.3 One Raw Dependent Variable and Five Raw Independent Variables

For the first hypothesis, all five variables are considered, and the results of the multiple regression are provided in Appendix 7 (yellow highlighted). The p-value of 0.596 for the age coefficient indicates that there is not enough evidence to support that β_5 is not equal to 0. So, it will be removed from the revised model. The other variables have p-values < 0.05, indicating that there is sufficient evidence to indicate that their associated coefficients are not 0. So, they will be kept in the revised model.

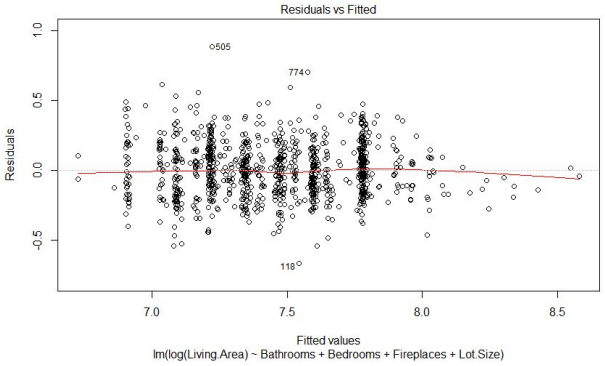


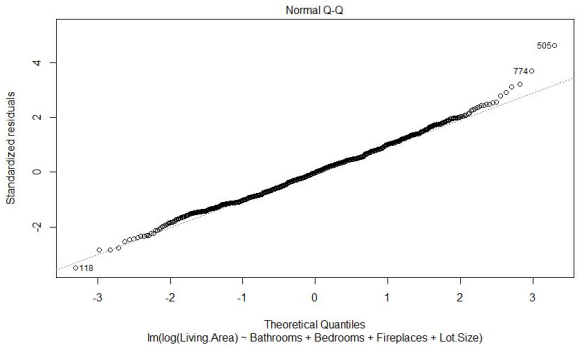


The scatter plot of fitted values vs. residuals above shows a slight curvilinear relationship, indicating that a linear model may not be the best fit for this hypothesis. The plot also shows that the equal variances assumption at all of the fitted values is violated, with the variance greater in the center. So, a transformation and/or introduction of a quadratic independent variable may help. The normal plot indicates that the residuals are not normally distributed, and this is confirmed by a Shapiro-Wilk test. The VIF values are all less than 3, so at least the multicollinearity assumption is not violated.

3.4 Log Transformed Dependent Variable and Four Raw Independent Variables

In the revised model, the living area is log transformed and the age variable is removed. The results of the multiple regression are provided in Appendix 7 (light blue highlighted). All of the variables have p-values < 0.05, indicating that there is sufficient evidence to indicate that their associated coefficients are not 0. Also, the adjusted R^2 increased from 0.6861 to 0.7066, indicating that this revised model better explains the variation in living area than the original model, without overfitting.





The scatter plot of fitted values vs. residuals above shows a linear relationship, indicating that the relationship between the independent and dependent variables is linear. The plot also shows an improvement in the equal variances assumption at all of the fitted values. The normal

plot isn't clear about the normality of the residuals, but it does show an improvement in normality over the original model. It appears to still be slightly non-normal due to the tails, and a Shapiro-Wilk test indicates that residuals are not normally distributed. The VIF values are again all less than 3, so the multicollinearity assumption is not violated.

Overall, the revised model is acceptable. Removal of certain outliers could help improve the normality of the distribution of the residuals, but then the model would less accurately represent the population. Use of the log transformation on the living area (dependent variable) makes the interpretation of the effects of the independent variables on the raw living area more difficult. For example, the bathroom coefficient of 0.247524 means that for each increase of 1 bathroom (and other variables held constant), the log(living area) increases by 0.247524. How does this translate to the living area? It turns out that for each increase of 1 for independent variable i (and other variables held constant) living area increases by $100 * (e^{\beta_i} - 1)$. The following table provides these values for each variable.

Variable :	Bathrooms	Bedrooms	Fireplaces	Lot Size
β :	0.247524	0.179398	0.133410	0.030316
$100 * (e^{\beta} - 1)$:	28.085%	19.650%	14.272%	3.078%

So, for each increase of 1 bathroom (and other variables held constant) the living area increases by 28.085%. The same argument applies to the other variables. When all independent variables are 0, the regression equation becomes $log(living\ area) = 6.294285$ which means that $living\ area = e^{6.294285} = 541.47$. So, a house with 0 bathrooms, 0 bedrooms, 0 fireplaces and 0 lot size has a living are of 541.47. The intercept really has no practical meaning, since a house can't be built on a lot with 0 size. Also, there are no actual observations with these values, so the intercept is an extrapolated value.

Appendix 1 - Data Sample

	A	B	C	D	E	F	G
1	Price	Living.Area	Bathrooms	Bedrooms	Fireplaces	Lot.Size	Age
2	142212	1982	1	3	0	2	133
3	134865	1676	1.5	3	1	0.38	14
4	118007	1694	2	3	1	0.96	15
5	138297	1800	1	2	2	0.48	49
6	129470	2088	1	3	1	1.84	29
7	206512	1456	2	3	0	0.98	10
8	108794	1464	1	2	0	0.11	87
9	68353	1216	1	2	0	0.61	101
10	123266	1632	1.5	3	0	0.23	14
11	309808	2270	2.5	3	2	4.05	9
12	157946	1804	2.5	3	1	0.43	0
13	80248	1600	1.5	3	0	0.36	16
14	135708	1460	2	2	0	0.18	17
15	173723	1548	2	3	1	0.36	0
16	140510	1590	2.5	3	1	0.42	0
17	122221	1170	1.5	4	0	3	26
18	151917	1510	2.5	3	1	0.39	0
19	235105	2299	2.5	4	1	0.8	6
20	259999	2577	2.5	4	1	0.77	1
21	211517	2328	2.5	4	1	0.85	10
22	102068	1172	2.5	3	1	0.85	73
23	128440	1554	1.5	3	0	4.87	103
24	115659	1242	2	3	1	0.72	30
25	145583	1376	2	3	1	0.46	25
26	116289	1107	1	3	1	0.46	43
27	238792	2250	2.5	4	1	2.48	10
28	221925	2472	2	4	0	0.62	183

Appendix 2 - Hypothesis 1 R Code

```
options(scipen=999)
priceage = read.table("C:\\Users\\Nick\\Desktop\\AMS\\572\\Project\\AMS 572 Housing
Data.csv", header=TRUE, sep=",",
colClasses=c(NA,"NULL","NULL","NULL","NULL","NULL",NA))
median(priceage$Age)

underequal18 = subset(priceage, Age<=18)[,1]
greater18 = subset(priceage, Age>18)[,1]
length(underequal18)
length(greater18)

boxplot(underequal18,greater18, names=c("Age<=18","Age>18"), ylab="Price")

shapiro.test(underequal18)
shapiro.test(greater18)

var.test(underequal18, greater18)

t.test(underequal18, greater18)

plot(priceage[,2],priceage[,1],xlim=c(0, 250), ylim=c(0, 600000),xlab="Age",ylab="Price")
```

Appendix 3 - Hypothesis 1 R Output

```
> options(scipen=999)
> priceage = read.table("C:\\Users\\Nick\\Desktop\\AMS\\572\\Project\\AMS 572 Housing
Data.csv", header=TRUE, sep=",",
colClasses=c(NA,"NULL","NULL","NULL","NULL","NULL",NA))
> median(priceage$Age)
[1] 18
> underequal18 = subset(priceage, Age<=18)[,1]
> greater18 = subset(priceage, Age>18)[,1]
> length(underequal18)
[1] 564
> length(greater18)
[1] 480
> boxplot(underequal18,greater18, names=c("Age<=18","Age>18"), ylab="Price")
> shapiro.test(underequal18)

        Shapiro-Wilk normality test

data:  underequal18
W = 0.93494, p-value = 0.000000000000005793

> shapiro.test(greater18)

        Shapiro-Wilk normality test

data:  greater18
W = 0.80714, p-value < 0.00000000000000022

> var.test(underequal18, greater18)

        F test to compare two variances

data:  underequal18 and greater18
F = 1.3731, num df = 563, denom df = 479, p-value = 0.0003433
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.154862 1.630611
```

```
sample estimates:
ratio of variances
         1.373106

> t.test(underequal18, greater18)

        Welch Two Sample t-test

data:  underequal18 and greater18
t = 12.431, df = 1042, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 45697.11 62828.03
sample estimates:
mean of x mean of y
 192516.8  138254.3

> plot(priceage[,2],priceage[,1],xlim=c(0, 250), ylim=c(0, 600000),xlab="Age",ylab="Price")
```

Appendix 4 - Hypothesis 2 R Code

```
library(car)
options(scipen=999)
fireplaceage = read.table("C:\\Users\\nhand\\Desktop\\AMS572Project\\AMS 572 Housing
Data.csv",
                    header=TRUE, sep=",",
colClasses=c("NULL","NULL","NULL","NULL",NA,"NULL",NA))
length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==0])
length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==1])
length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==2])
length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==3])
length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==4])
length(fireplaceage$Fireplaces[fireplaceage$Fireplaces>=5])

fireplaceage$Fireplaces[fireplaceage$Fireplaces==3] = 2
fireplaceage$Fireplaces[fireplaceage$Fireplaces==4] = 2

fireplaceage$Fireplaces[fireplaceage$Fireplaces==0] = '0'
fireplaceage$Fireplaces[fireplaceage$Fireplaces==1] = '1'
fireplaceage$Fireplaces[fireplaceage$Fireplaces==2] = '2'

#raw data analysis
boxplot(fireplaceage$Age~fireplaceage$Fireplaces, names=c("0","1","2+"), ylab="Age",
xlab="Fireplaces")

shapiro.test(fireplaceage$Age[fireplaceage$Fireplaces=='0'])
shapiro.test(fireplaceage$Age[fireplaceage$Fireplaces=='1'])
shapiro.test(fireplaceage$Age[fireplaceage$Fireplaces=='2'])
leveneTest(fireplaceage$Age~fireplaceage$Fireplaces)

fit = aov(fireplaceage$Age~fireplaceage$Fireplaces)
summary(fit)
plot(fit)

shapiro.test(fit$residuals)

#log(x+1) transformed data analysis
```

```
fireplacelogage = fireplaceage
fireplacelogage$Age = log(fireplacelogage$Age+1)

boxplot(fireplacelogage$Age~fireplacelogage$Fireplaces, names=c("0","1","2+"),
ylab="log(Age+1)", xlab="Fireplaces")

shapiro.test(fireplacelogage$Age[fireplacelogage$Fireplaces=='0'])
shapiro.test(fireplacelogage$Age[fireplacelogage$Fireplaces=='1'])
shapiro.test(fireplacelogage$Age[fireplacelogage$Fireplaces=='2'])
leveneTest(fireplacelogage$Age~fireplacelogage$Fireplaces)

fit = aov(fireplacelogage$Age~fireplacelogage$Fireplaces)
summary(fit)
plot(fit)

shapiro.test(fit$residuals)
```

Appendix 5 - Hypothesis 2 R Output

```
> library(car)
> options(scipen=999)
> fireplaceage = read.table("C:\\Users\\nhand\\Desktop\\AMS572Project\\AMS 572 Housing
Data.csv",
+                           header=TRUE, sep=",",
colClasses=c("NULL","NULL","NULL","NULL",NA,"NULL",NA))
> length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==0])
[1] 422
> length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==1])
[1] 595
> length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==2])
[1] 24
> length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==3])
[1] 2
> length(fireplaceage$Fireplaces[fireplaceage$Fireplaces==4])
[1] 1
> length(fireplaceage$Fireplaces[fireplaceage$Fireplaces>=5])
[1] 0
> fireplaceage$Fireplaces[fireplaceage$Fireplaces==3] = 2
> fireplaceage$Fireplaces[fireplaceage$Fireplaces==4] = 2
> fireplaceage$Fireplaces[fireplaceage$Fireplaces==0] = '0'
> fireplaceage$Fireplaces[fireplaceage$Fireplaces==1] = '1'
> fireplaceage$Fireplaces[fireplaceage$Fireplaces==2] = '2'
> #raw data analysis
> boxplot(fireplaceage$Age~fireplaceage$Fireplaces, names=c("0","1","2+"), ylab="Age",
xlab="Fireplaces")
> shapiro.test(fireplaceage$Age[fireplaceage$Fireplaces=='0'])

        Shapiro-Wilk normality test

data:  fireplaceage$Age[fireplaceage$Fireplaces == "0"]
W = 0.79249, p-value < 0.00000000000000022

> shapiro.test(fireplaceage$Age[fireplaceage$Fireplaces=='1'])

        Shapiro-Wilk normality test
```

```
data:  fireplaceage$Age[fireplaceage$Fireplaces == "1"]
W = 0.66163, p-value < 0.00000000000000022

> shapiro.test(fireplaceage$Age[fireplaceage$Fireplaces=='2'])

        Shapiro-Wilk normality test

data:  fireplaceage$Age[fireplaceage$Fireplaces == "2"]
W = 0.59254, p-value = 0.0000001692

> leveneTest(fireplaceage$Age~fireplaceage$Fireplaces)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value        Pr(>F)
group   2   19.63 0.000000004281 ***
      1041
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.
> fit = aov(fireplaceage$Age~fireplaceage$Fireplaces)
> summary(fit)
                            Df      Sum Sq   Mean Sq   F value            Pr(>F)
fireplaceage$Fireplaces   2       85711     42855      37.23       0.000000000000000242 ***
Residuals                 1041   1198404    1151
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> plot(fit)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.76278, p-value < 0.00000000000000022
```

```
> #log(x+1) transformed data analysis
> fireplacelogage = fireplaceage
> fireplacelogage$Age = log(fireplacelogage$Age+1)
> boxplot(fireplacelogage$Age~fireplacelogage$Fireplaces, names=c("0","1","2+"),
ylab="log(Age+1)", xlab="Fireplaces")
> shapiro.test(fireplacelogage$Age[fireplacelogage$Fireplaces=='0'])

        Shapiro-Wilk normality test

data:  fireplacelogage$Age[fireplacelogage$Fireplaces == "0"]
W = 0.95812, p-value = 0.000000001369

> shapiro.test(fireplacelogage$Age[fireplacelogage$Fireplaces=='1'])

        Shapiro-Wilk normality test

data:  fireplacelogage$Age[fireplacelogage$Fireplaces == "1"]
W = 0.92766, p-value = 0.0000000000000002476

> shapiro.test(fireplacelogage$Age[fireplacelogage$Fireplaces=='2'])

        Shapiro-Wilk normality test

data:  fireplacelogage$Age[fireplacelogage$Fireplaces == "2"]
W = 0.94725, p-value = 0.1837

> leveneTest(fireplacelogage$Age~fireplacelogage$Fireplaces)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   2  7.9149 0.0003877 ***
      1041
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.
> fit = aov(fireplacelogage$Age~fireplacelogage$Fireplaces)
> summary(fit)
                          Df Sum Sq Mean Sq F value             Pr(>F)
fireplacelogage$Fireplaces    2  159.2   79.60   50.09 <0.0000000000000002 ***
```

```
Residuals                1041 1654.2    1.59
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> plot(fit)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.967, p-value = 0.00000000000001197
```

Appendix 6 - Hypothesis 3 R Code

```
library(car)
options(scipen=999)

housedata = read.table("C:\\Users\\nhand\\Desktop\\AMS572Project\\AMS 572 Housing
Data.csv",
                       header=TRUE, sep=",")

plot(housedata$Bathrooms, housedata$Living.Area)
plot(housedata$Bedrooms, housedata$Living.Area)
plot(housedata$Fireplaces, housedata$Living.Area)
plot(housedata$Lot.Size, housedata$Living.Area)
plot(housedata$Age, housedata$Living.Area)

fit <- lm(Living.Area ~ Bathrooms+Bedrooms+Fireplaces+Lot.Size+Age, data=housedata)
summary(fit)
plot(fit)
shapiro.test(fit$residuals)
vif(fit)

fit <- lm(log(Living.Area) ~ Bathrooms+Bedrooms+Fireplaces+Lot.Size, data=housedata)
summary(fit)
plot(fit)
shapiro.test(fit$residuals)
vif(fit)
```

Appendix 7 - Hypothesis 3 R Output

```
> library(car)
> options(scipen=999)
> housedata = read.table("C:\\Users\\nhand\\Desktop\\AMS572Project\\AMS 572 Housing
Data.csv",
+                        header=TRUE, sep=",")
> plot(housedata$Bathrooms, housedata$Living.Area)
> plot(housedata$Bedrooms, housedata$Living.Area)
> plot(housedata$Fireplaces, housedata$Living.Area)
> plot(housedata$Lot.Size, housedata$Living.Area)
> plot(housedata$Age, housedata$Living.Area)
> fit <- lm(Living.Area ~ Bathrooms+Bedrooms+Fireplaces+Lot.Size+Age, data=housedata)
> summary(fit)

Call:
lm(formula = Living.Area ~ Bathrooms + Bedrooms + Fireplaces +
    Lot.Size + Age, data = housedata)

Residuals:
     Min       1Q   Median       3Q      Max
-1046.89  -249.05   -34.28   193.36  1918.77

Coefficients:
             Estimate   Std. Error  t value            Pr(>|t|)
(Intercept)  -306.2138     55.5406   -5.513       0.0000000444 ***
Bathrooms     465.6220     24.2734   19.182  < 0.0000000000000002 ***
Bedrooms      324.9219     18.6468   17.425  < 0.0000000000000002 ***
Fireplaces    240.8739     23.3142   10.332  < 0.0000000000000002 ***
Lot.Size       65.3176     15.1304    4.317       0.0000173311 ***
Age             0.1962      0.3698    0.530              0.596
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 371.3 on 1038 degrees of freedom
Multiple R-squared:  0.6876,  Adjusted R-squared:  0.6861
F-statistic:   457 on 5 and 1038 DF,  p-value: < 0.00000000000000022
```

```
> plot(fit)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.96445, p-value = 0.000000000000002808

> vif(fit)
 Bathrooms   Bedrooms Fireplaces   Lot.Size        Age
  1.893544   1.429645   1.247962   1.018636   1.273380
> fit <- lm(log(Living.Area) ~ Bathrooms+Bedrooms+Fireplaces+Lot.Size, data=housedata)
> summary(fit)

Call:
lm(formula = log(Living.Area) ~ Bathrooms + Bedrooms + Fireplaces +
    Lot.Size, data = housedata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66911 -0.13830 -0.00228  0.11476  0.88786

Coefficients:
             Estimate   Std. Error    t value          Pr(>|t|)
(Intercept)  6.294285    0.027061    232.599          < 0.0000000000000002 ***
Bathrooms    0.247524    0.011401    21.712           < 0.0000000000000002 ***
Bedrooms     0.179398    0.009455    18.974           < 0.0000000000000002 ***
Fireplaces   0.133410    0.012041    11.079           < 0.0000000000000002 ***
Lot.Size     0.030316    0.007818    3.878            0.000112 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1921 on 1039 degrees of freedom
Multiple R-squared:  0.7077,  Adjusted R-squared:  0.7066
F-statistic: 628.9 on 4 and 1039 DF,  p-value: < 0.00000000000000022
```

```
> plot(fit)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.99483, p-value = 0.001223

> vif(fit)
 Bathrooms   Bedrooms  Fireplaces   Lot.Size
  1.561145   1.373834    1.244168   1.016564
```