STAT 8030 Homework 3

Connor Armstrong

12/13/2020

Attaching necessary libraries

```r
library(tidyverse)#data manipulation
library(ggplot2)#plots
library(dplyr)#data manipulation
library(equatiomatic)#pretty equations
library(RVAideMemoire)#moods median test, not used
library(car)#anova
library(nortest)#normality testing
library(ggpubr)#plots
library(knitr)#tables
```

PART I

Use the ggplot2::txhousing dataframe to perform some statistical analysis. Specifically:

1. Subset the dataframe such that we're only looking at homes in Austin, Texas (for obvious reasons haha) between (and inclusive of) 2010 - 2015.

```r
house <- ggplot2::txhousing

house1 <- house %>%
  filter(city == "Austin" & year > 2009 & year < 2016)
```

2. Suppose I want to compare the median home sale prices in Austin across those six years. Treating year as a categorical variable, build an appropriate statistical model. Be sure to write out what the equation is. (HINT: use equatiomatic).

One way ANOVA

```r
res.aov <- aov(median ~ year, data = house1)
```

Summary of the analysis

```r
summary(res.aov)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## year         1 3.546e+10 3.546e+10   278.3 <2e-16 ***
## Residuals   65 8.283e+09 1.274e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value is significant ($p < 0.05$), there is a statistically significant difference of median prices of homes between some of the years in the data set.

What is the mean and standard deviation of the median prices by year?

```
med_sum <- group_by(house1, year) %>%
  summarise(
    count = n(),
    mean = mean(median, na.rm = TRUE),
    sd = sd(median, na.rm = TRUE)
  )


kable(med_sum, caption = ’Summary of Median Home Prices in Austin, TX’)
```

Table 1: Summary of Median Home Prices in Austin, TX

year	count	mean	sd
2010	12	189658.3	10457.573
2011	12	190033.3	4816.512
2012	12	201741.7	10714.600
2013	12	220508.3	9819.318
2014	12	238791.7	9580.421
2015	7	259000.0	13597.059

What does the equation for the ANOVA model look like?

$\text{median}_{ij} = \beta_0 + \beta_1 * (\text{year}) + \epsilon_{ij}$

$\text{median}_{ij} = -28317042 + 14178 * (\text{year}) + \epsilon_{ij}$

3. Assess all relevant assumptions

- Assessing homogeneity of variance

Levenes test needs year to be categorical:

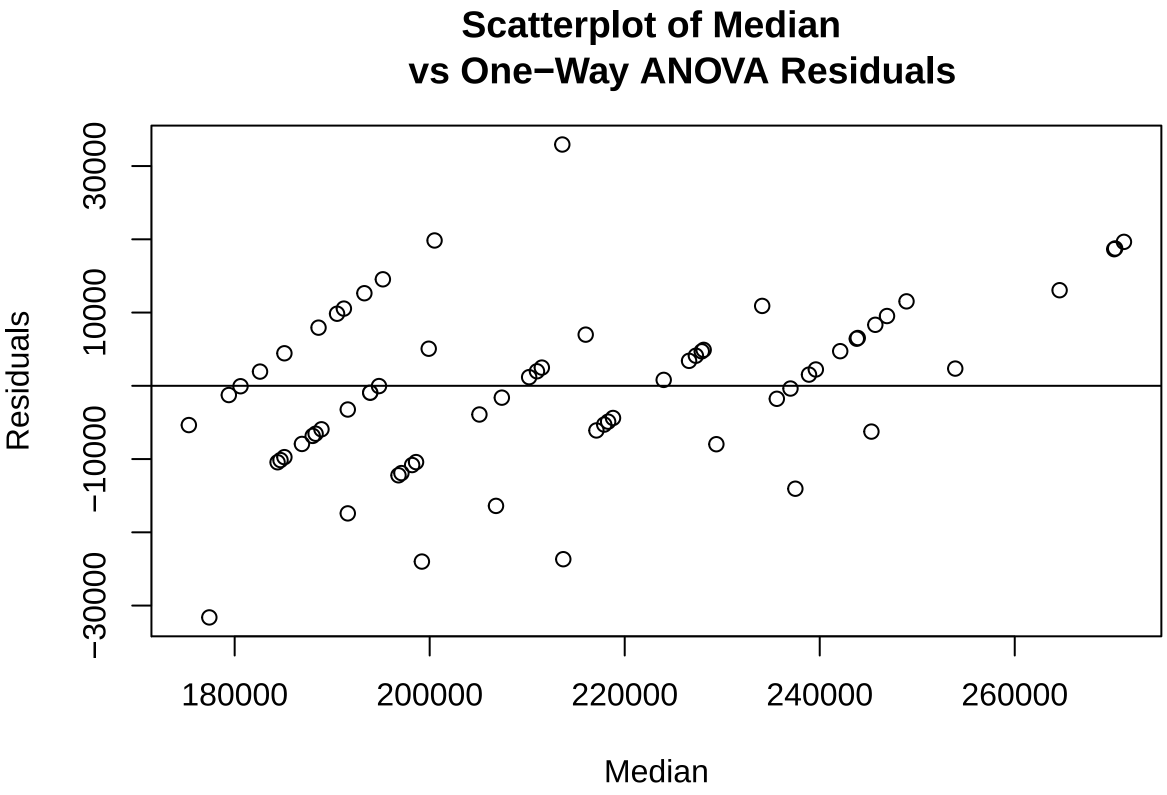
```
house1$yearcat <-cut(house1$year,
                     breaks = c(2009.5,2010.5,2011.5,2012.5,2013.5,2014.5, 2015.5),
                     labels=c("2010","2011","2012","2013","2014","2015"))

leveneTest(median ~ yearcat, data = house1)
```

```
## Levene’s Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  1.2554 0.2948
##       61
```

With a p value $0.29 > 0.05$, we can conclude that there is no evidence to suggest variance across groups is significantly different, validating the HOV assumption

- Are the residuals normal?

```
resid <- resid(res.aov)
plot(house1$median, resid, xlab="Median", ylab="Residuals", main = "Scatterplot of Median
     vs One-Way ANOVA Residuals")
abline(0,0)
```



```
ad.test(resid)
```

```
##
##  Anderson-Darling normality test
##
## data:  resid
## A = 0.25938, p-value = 0.7033
```

With a p-value of > 0.05 from the Anderson-Darling normality test, it is likely that the residuals are normally distributed.

Which days are different? Wilcox test with p adjust method BH recommended per

http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r

```
pairwise.wilcox.test(house1$median, house1$yearcat,
                     p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
```

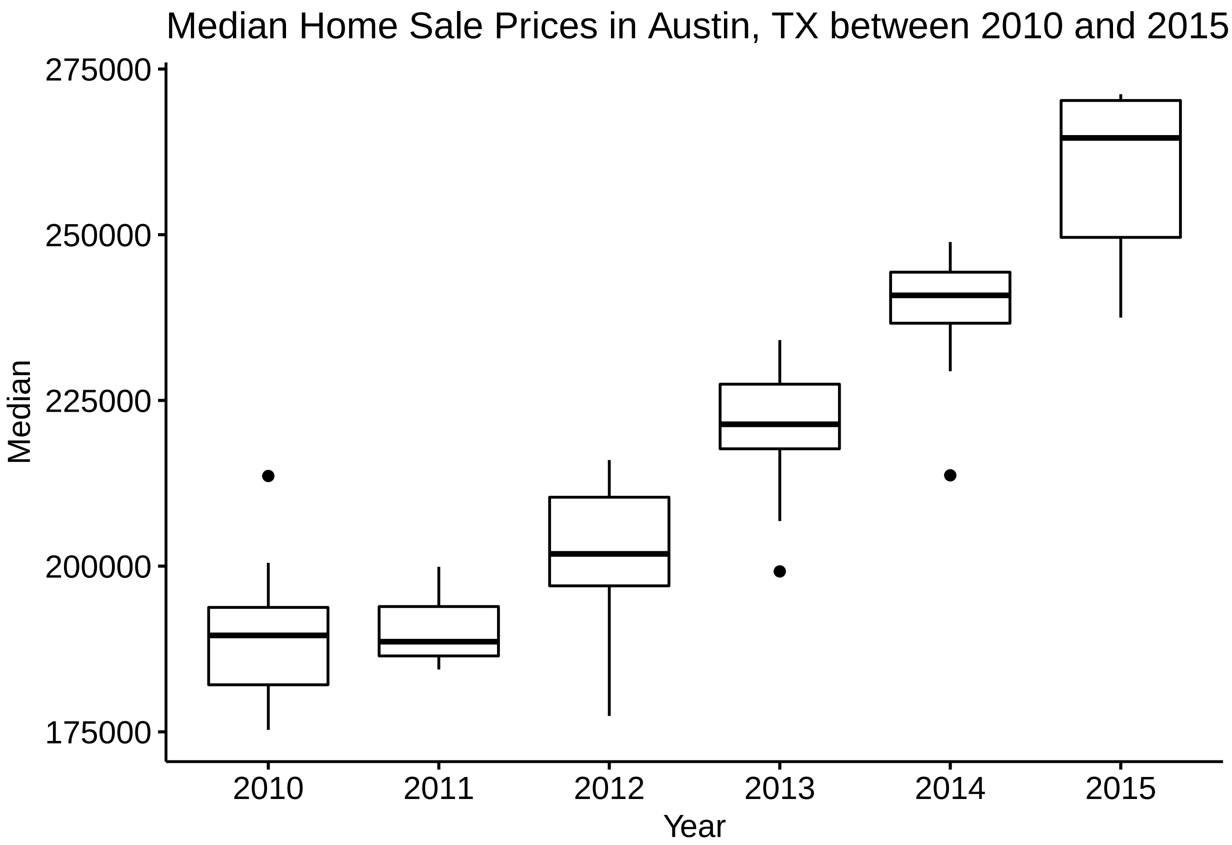
```
##
## data:  house1$median and house1$yearcat
##
##      2010    2011    2012    2013    2014
## 2011 0.72892 -       -       -       -
## 2012 0.01076 0.00403 -       -       -
## 2013 2.6e-05 8.8e-05 0.00022 -       -
## 2014 1.1e-05 8.5e-05 1.1e-05 0.00022 -
## 2015 8.5e-05 0.00061 8.5e-05 8.5e-05 0.00829
##
## P value adjustment method: BH
```

Combinations with p < 0.05 are significantly different. All combinations of years but 2010 & 2011 have p < 0.05 and are significantly different.

4. Using whichever method you like, generate a graphic as well as a table which contain the appropriate outputs.

Box and Stem plot by year of Median Home Sale Prices in Austin, TX between 2010 and 2015

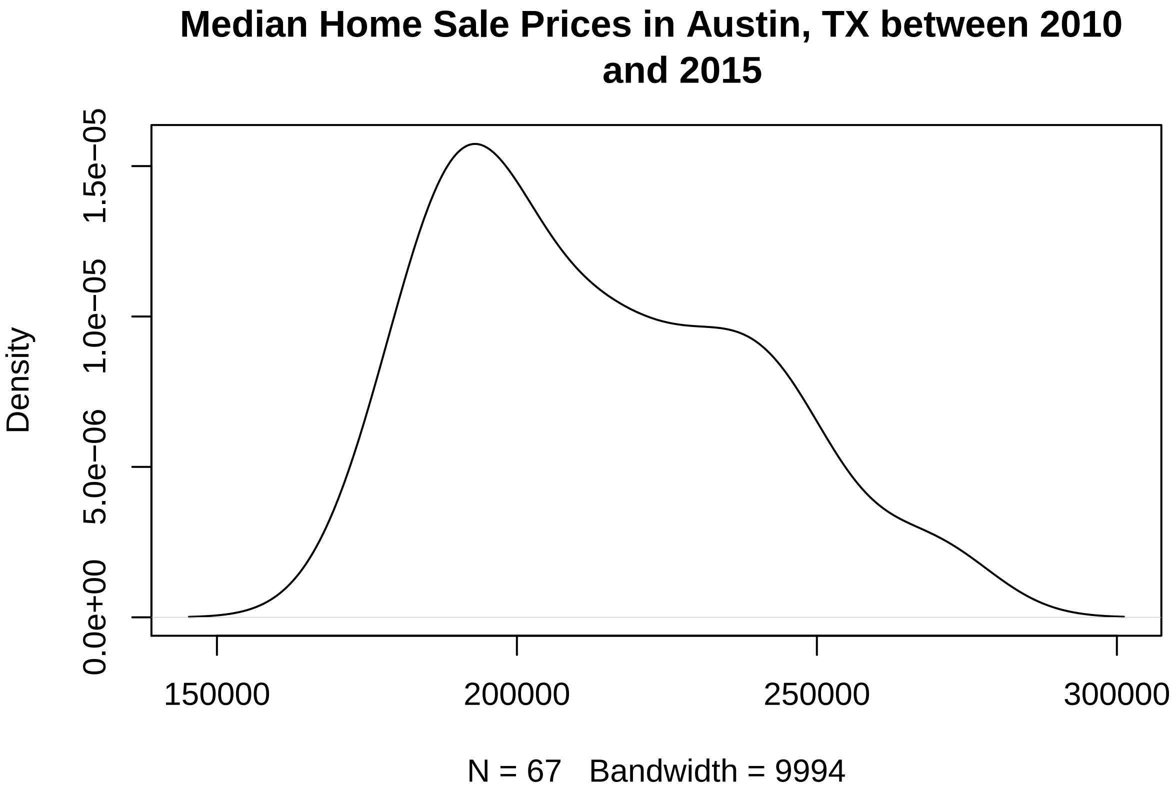
```
ggboxplot(house1, x = 'year', y = 'median',
          ylab = "Median", xlab = "Year",
          title = "Median Home Sale Prices in Austin, TX between 2010 and 2015")
```



There appears to be a steady increase in median prices by year, while the variation in median prices does not appear to be changing significantly, upon visual inspection. These conclusions agree with the results of the ANOVA (significant differences between the mean medians between years) and Levene's test of HOV (no significant differences in variance between years).

Density plot of Median Home Sale Prices in Austin, TX between 2010 and 2015

```
plot(density(house1$median), main = "Median Home Sale Prices in Austin, TX between 2010
    and 2015")
```



Visual inspection of the density plot indicates a potential bimodal distribution. This is indicative of some external influence to the population which causes deviation from normal behavior.

5. In the context of the problem, state what inference can be made.

The median home prices for the 6 years are significantly different from each other, except for 2010 and 2011. Each year after 2011 has a mean median sale price greater than the last, therefore the data suggest that since 2011 median home prices have risen significantly every year.

PART II

For part two, use the same dataframe you created in part one for a new analysis. Here, suppose I have an inclination that median sale price and number of sales are interrelated. However, I also suspect that the quarter of the year in which the sale occured also has an effect of median home sale price. So specifically:

1. Create a new variable called "quarter" where months 1-3 are "Q1", months 4-6 are "Q2", months 7-9 are "Q3", and months 10-12 are "Q4".

```
house1$quarter <-cut(house1$month,
                     breaks = c(0,3.5,6.5,9.5,12.5),
                     labels=c("Q1","Q2","Q3","Q4"))
```

2. Build a model where median is the outcome variable, sales is the first explanatory variable, and quarter is the second explanatory variable. Be sure to write out what the equation is. (HINT: use equatiomatic).

```
pt2 <- lm(median ~ sales + quarter, data = house1)
summary(pt2)
```

```
##
## Call:
## lm(formula = median ~ sales + quarter, data = house1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36528  -9601  -1506   7291  40558
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 140178.296   7825.658  17.913  < 2e-16 ***
## sales           38.123      3.981   9.576 7.73e-14 ***
## quarterQ2   -17218.626   6531.395  -2.636  0.01058 *
## quarterQ3   -18992.579   6339.969  -2.996  0.00393 **
## quarterQ4    -5023.606   5716.813  -0.879  0.38293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16190 on 62 degrees of freedom
## Multiple R-squared:  0.6285, Adjusted R-squared:  0.6045
## F-statistic: 26.22 on 4 and 62 DF,  p-value: 9.559e-13
```

The model parameters from the simple linear regression summary indicate that the chosen explanatory variables appear to have some relationship with the outcome variable. This will need to be explored in more detail to determine whether that relationship exists, and whether the appropriate model assumptions are met to reach such a conclusion.

What does the equation for the simple linear regression model look like?

```
#Symbols
extract_eq(pt2)
```

$$\operatorname{median} = \alpha + \beta_{1}(\operatorname{sales}) + \beta_{2}(\operatorname{quarter}_{\operatorname{Q2}}) + \beta_{3}(\operatorname{quarter}_{\operatorname{Q3}}) + \beta_{4}(\operatorname{quarter}_{\operatorname{Q4}}) + \epsilon$$

```
#Values
extract_eq(pt2, use_coefs = T, coef_digits = 1)
```

$$\text{median} = 140178.3 + 38.1(\text{sales}) - 17218.6(\text{quarter}_{\text{Q2}}) - 18992.6(\text{quarter}_{\text{Q3}}) - 5023.6(\text{quarter}_{\text{Q4}}) + \epsilon$$

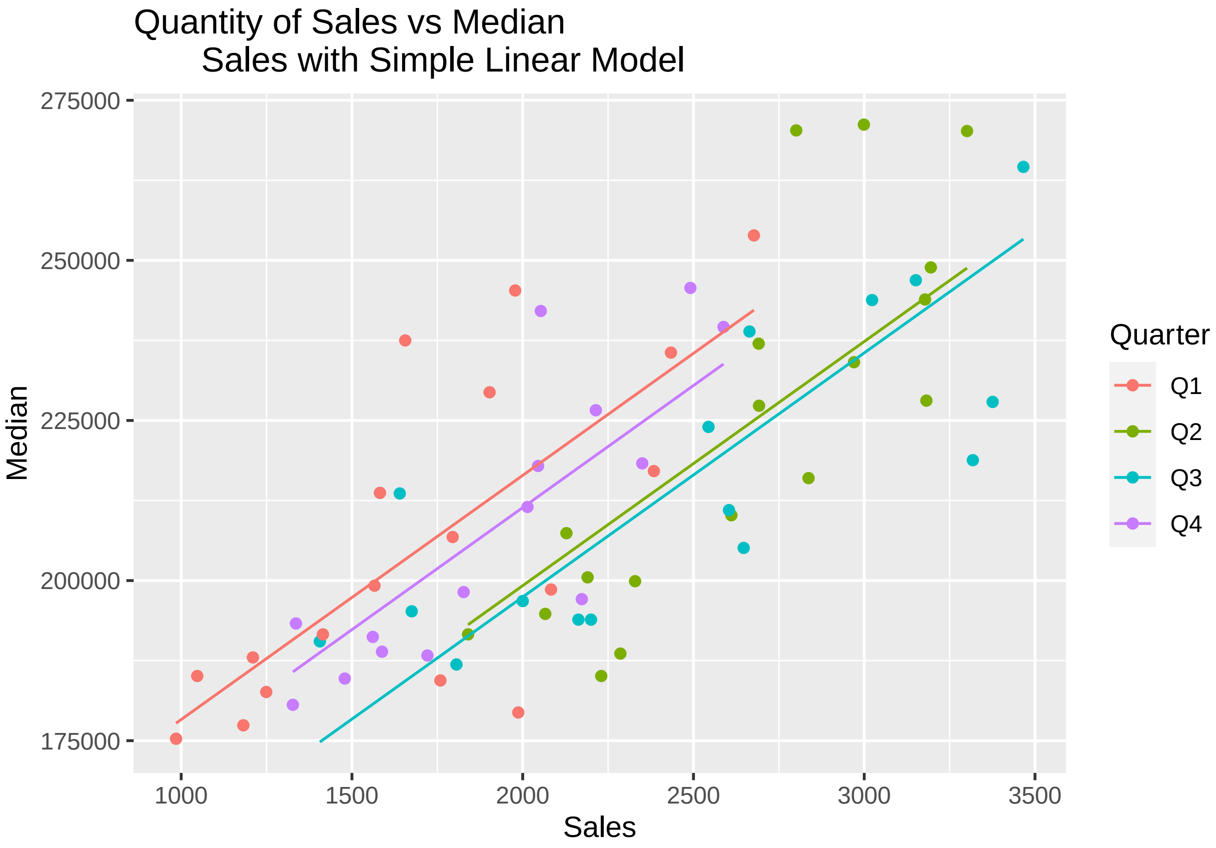
```
house1 <- cbind(house1, pred = predict(pt2))
```

The coefficients of this equation imply some interesting relationships between these variables. These implications must be interpreted contextually, according to the statistical significance of the coefficients, as well as how the coefficients relate to one another. The conclusions of any discussion related to this model are contingent on validation of the appropriate assumptions.

With that said, the following is a brief analysis of the coefficients of the above multiple linear regression equation.

The variable "Quarter" is a categorical variable, and is therefore dummy coded in the equation. At each of the 4 possible values of "Quarter", the regression line shifts vertically to place the regression line in the best fit location to minimize the errors (SRSS of the Residuals).

```
ggplot(house1, aes(x = sales, y = median, colour = quarter)) +
  geom_point() +
  geom_line(mapping = aes(y=pred)) +
  labs(x = "Sales", y = "Median", colour = "Quarter", title = "Quantity of Sales vs Median
       Sales with Simple Linear Model")
```



Interestingly, Q1 is not shown in the equation. This is done automatically in R to reduce the complexity of the model. Given that the coefficient for Q1 would be constant, it is automatically lumped with the intercept and the coefficients of Q2, Q3, and Q4 encode the difference of each from Q1.

A more complex model might consider the varying slope of the best fit line due to the relationship between sales and median by quarter, by combining the "Quarter" dummy variable with sales.

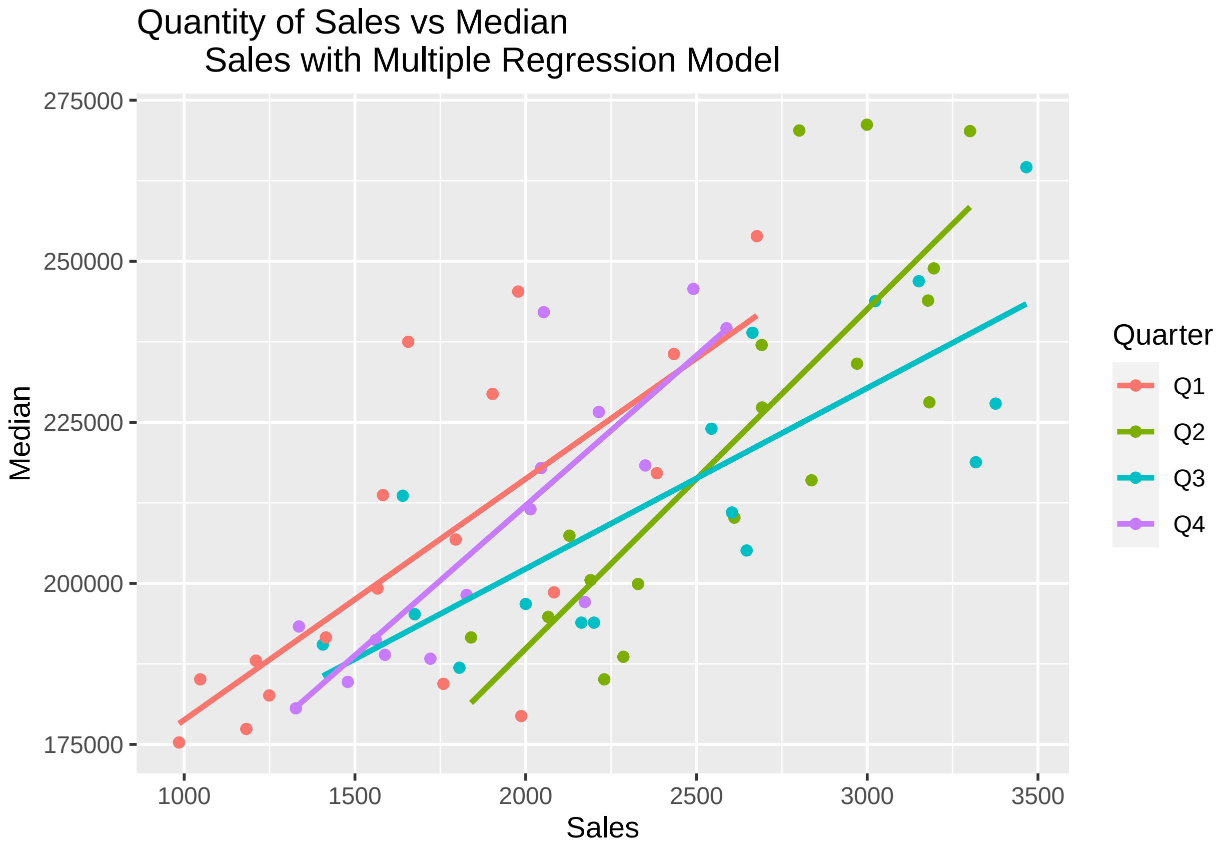
```
pt2_xterm <- lm(median ~ sales + quarter + sales*quarter, data = house1)
summary(pt2_xterm)
```

```
##
## Call:
## lm(formula = median ~ sales + quarter + sales * quarter, data = house1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36341 -10006   -783  10731  38216
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     141368.324  13903.134  10.168 1.37e-14 ***
## sales               37.429      7.806   4.795 1.14e-05 ***
## quarterQ2       -56906.242  26608.337  -2.139   0.0366 *
## quarterQ3         4799.256  20929.254   0.229   0.8194
## quarterQ4       -22379.676  24564.051  -0.911   0.3660
## sales:quarterQ2     15.274     11.523   1.325   0.1901
## sales:quarterQ3     -9.379      9.909  -0.946   0.3478
## sales:quarterQ4      9.123     12.959   0.704   0.4842
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 15770 on 59 degrees of freedom
## Multiple R-squared:  0.6646, Adjusted R-squared:  0.6248
## F-statistic:  16.7 on 7 and 59 DF,  p-value: 6.215e-12
```

Several of the coefficients are not significant, and the Rˆ2 value for this model is not significantly better than the simpler model. The standard error is only marginally smaller than for the simpler model. For these reasons, this model will not be discussed or analyzed further.

```
#x-term plot
ggplot(house1, aes(sales, median, colour = (quarter))) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  labs(x = "Sales", y = "Median", colour = "Quarter", title = "Quantity of Sales vs Median
       Sales with Multiple Regression Model")
```

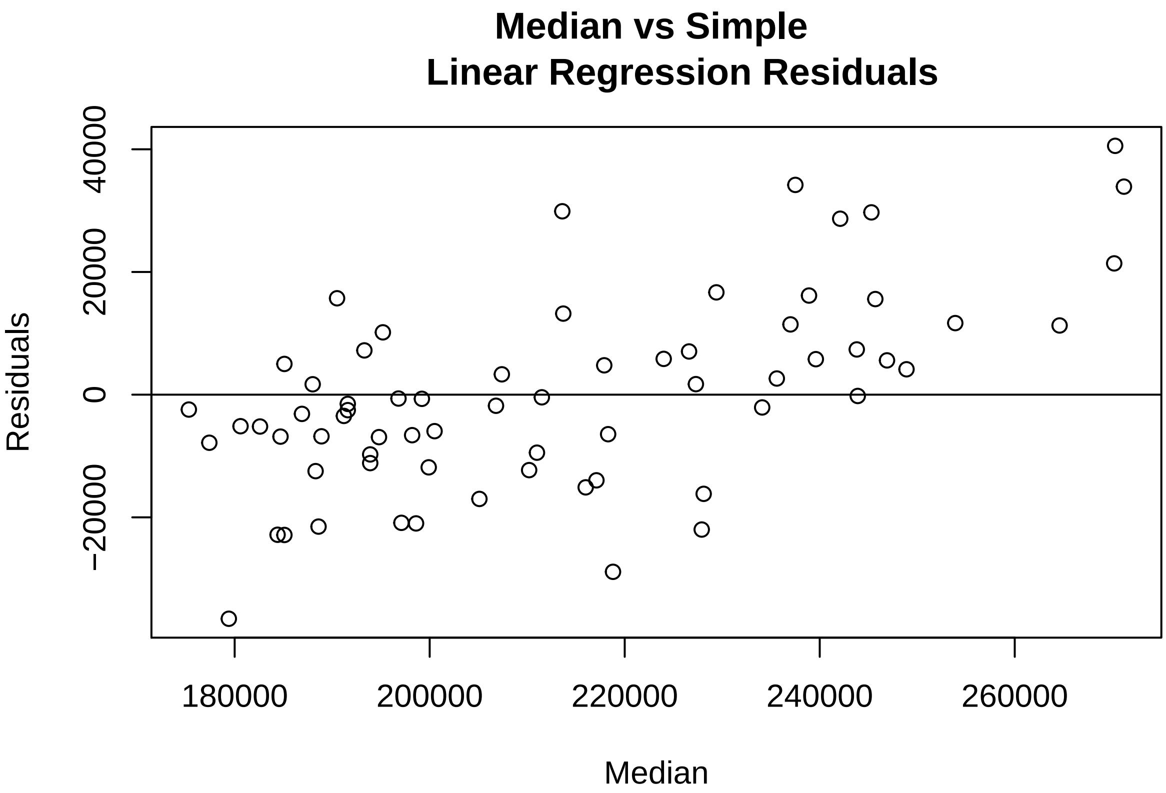
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



3. Assess all relevant assumptions.

- Are the residuals for the chosen model (without cross-terms) normally distributed?

```
resid2 <- resid(pt2)
plot(house1$median, resid2, xlab="Median", ylab="Residuals", main = "Median vs Simple
     Linear Regression Residuals")
abline(0,0)
```



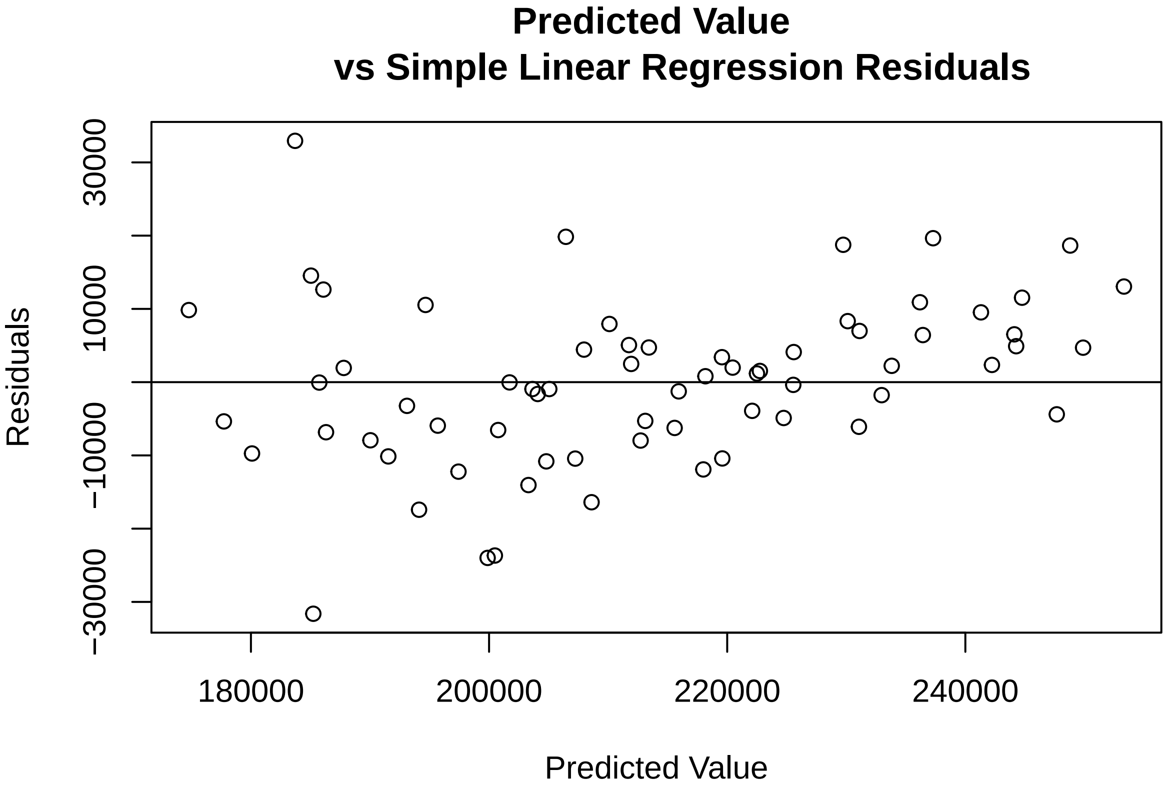
```r
ad.test(resid2)
```

```
##
##  Anderson-Darling normality test
##
## data:  resid2
## A = 0.45578, p-value = 0.2596
```

While the scatterplot of residuals vs the outcome variable is not perfectly random, the Anderson-Darling normality test returned a p-value of > 0.05, which meets the criteria for normality for this analysis.

- Checking the homogeneity of variance assumption.

```r
plot(house1$pred, resid, xlab="Predicted Value", ylab="Residuals", main = "Predicted Value
     vs Simple Linear Regression Residuals")
abline(0,0)
```



Visual inspection of the residuals vs predicted value plot does not present significant evidence of non-homogeneity of variance.

- Checking independence assumption with Durbin Watson test.

```
durbinWatsonTest(pt2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5977174     0.7963722       0
##  Alternative hypothesis: rho != 0
```

The Durbin Watson helps to determine if the errors in the model are autocorrelated. If p > 0.05, they likely are not. The test returned a p-value of 0, which implies the median price and number of sales are not independent. This is intuitively consistent with expected behavior, as higher demand (and therefore sales) often results in higher prices. This assumption violation unfortunately calls into question the predictive ability of the model and should not be ignored when drawing conclusions from the model results.

4. Using whichever method you like, generate a graphic as well as a table which contain the appropriate outputs.

See scatterplots and residual plots above.

The following is an ANOVA and simple table comparing the diagnostics for the two regression models.

```
anova(pt2, pt2_xterm)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: median ~ sales + quarter
## Model 2: median ~ sales + quarter + sales * quarter
##   Res.Df        RSS Df  Sum of Sq     F Pr(>F)
## 1     62 1.6252e+10
## 2     59 1.4672e+10  3 1580056895 2.118 0.1076
```

```
x <- tribble(
  ~Name,~r2,~Error,
  "Linear",summary(pt2)$r.squared,sqrt(deviance(pt2)/df.residual(pt2)),
  "Multiple",summary(pt2_xterm)$r.squared,sqrt(deviance(pt2_xterm)/df.residual(pt2_xterm))
)

kable(x, caption = 'Comparison of Model Diagnostics')
```

Table 2: Comparison of Model Diagnostics

Name	r2	Error
Linear	0.6284915	16190.4
Multiple	0.6646104	15769.5

The ANOVA returned a p-value > 0.05, which inicates that the performance of the more complex regression model did not perform significantly different from the simple model.

5. In the context of the problem, state what inference can be made.

There is certainly some degree of a linear relationship between sales and median home prices in this data set when accounting for differences between seasons given the r^2 value of the chosen model implied that ~60% of the variability in median home prices can be explained by the chosen explanatory variables. As mentioned above, the violation of the independence assumption calls into question the credibility of any conclusions one might come to when implementing this type of model.