Project One: Multiple Regression, Qualitative Variables Interactions, Quadratic Regression

Scenario

A data analyst working for a real estate company has access to a large set of historical data that is to be used to analyze relationships between different attributes of a house (such as square footage or the number of bathrooms) and the house's selling price.

Different regression models were created to predict sale prices for houses based on critical variable factors. These regression models will help the real estate company set better prices when listing a home for a client. Setting better prices will ensure that listings can be sold within a reasonable amount of time.

These important variables are used in the modeling:

Variable	**What does it represent?**
price	Sale price of the home
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
sqft_living	Size of the living area in sqft
sqft_above	Size of the upper level in sqft
sqft_lot	Size of the lot in sqft
age	Age of the home
grade	Measure of craftsmanship and the quality of materials used to build the home
appliance_age	Average age of all appliances in the home
crime	Crime rate per 100,000 people
backyard	Home has a backyard (backyard=1) or not (backyard=0)
school_rating	Average rating of schools in the area
view	Home backs out to a lake (view=2), backs out to trees (view=1), or backs out to a road (view=0)

R code is used in a Jupyter Notebook environment

Data set preparation:

```r
housing <- read.csv(file="housing_v2.csv", header=TRUE, sep=",")

# converting appropriate variables to factors
housing <- within(housing, {
   view <- factor(view)
   backyard <- factor(backyard)
})

# number of columns
ncol(housing)

# number of rows
nrow(housing)
```

23

2692

Model #1 - First Order Regression Model with Quantitative and Qualitative Variables

A first order regression model is created for price as the response variable, and sqft_living, sqft_above, age, bathrooms, and view as predictor variables.

```r
housing <- read.csv(file="housing_v2.csv", header=TRUE, sep=",")

# converting appropriate variables to factors
housing <- within(housing, {
   view <- factor(view)
   backyard <- factor(backyard)
})

plot(housing$sqft_living, housing$price,
     main = "Scatterplot of price against living area in sq ft",
     xlab = "living area in sq ft", ylab = "price",
     col="red",
     pch = 19, frame = FALSE)
```

```r
housing <- read.csv(file="housing_v2.csv", header=TRUE, sep=",")

# converting appropriate variables to factors
housing <- within(housing, {
   view <- factor(view)
   backyard <- factor(backyard)
})

plot(housing$sqft_above, housing$price,
     main = "Scatterplot of price against upper living area in sq ft",
     xlab = "upper living area in sq ft", ylab = "price",
     col="red",
     pch = 19, frame = FALSE)
```

```
myvars <- c("price","sqft_living")
housing_subset <- housing[myvars]

# Print correlation matrix
print("cor")
corr_matrix <- cor(housing_subset, method = "pearson")
round(corr_matrix, 4)
```

```
[1] "cor"
```

A matrix: 2 × 2 of type dbl

	price	sqft_living
price	1.0000	0.6895
sqft_living	0.6895	1.0000

```
myvars <- c("price","age")
housing_subset <- housing[myvars]

# Print correlation matrix
print("cor")
corr_matrix <- cor(housing_subset, method = "pearson")
round(corr_matrix, 4)
```

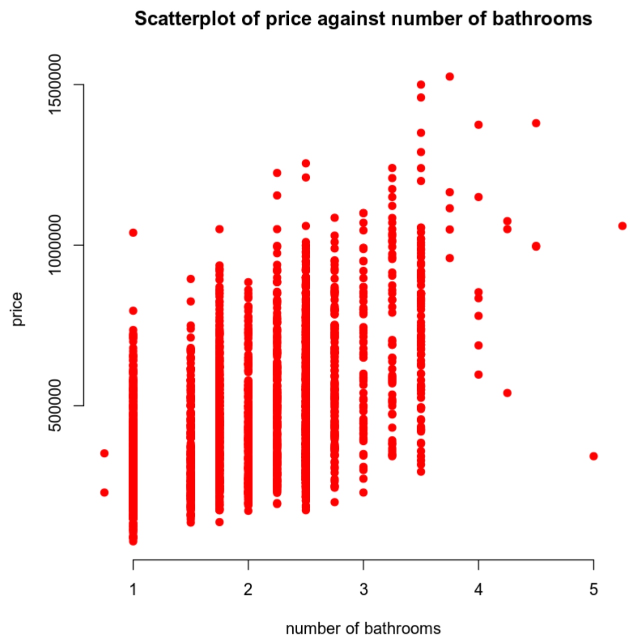
```
[1] "cor"
```

A matrix: 2 × 2 of type
dbl

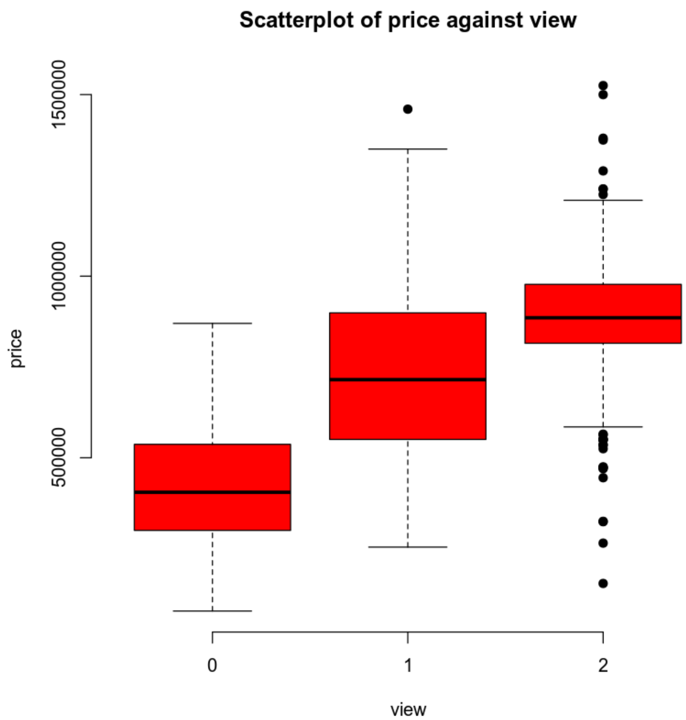
	price	age
price	1.0000	-0.0746
age	-0.0746	1.0000

```
housing <- read.csv(file="housing_v2.csv", header=TRUE, sep=",")

# converting appropriate variables to factors
housing <- within(housing, {
   view <- factor(view)
   backyard <- factor(backyard)
})

plot(housing$age, housing$price,
     main = "Scatterplot of price against age of home",
     xlab = "age of home in years", ylab = "price",
     col="red",
     pch = 19, frame = FALSE)
```


```r
housing <- read.csv(file="housing_v2.csv", header=TRUE, sep=",")

# converting appropriate variables to factors
housing <- within(housing, {
   view <- factor(view)
   backyard <- factor(backyard)
})

plot(housing$bathrooms, housing$price,
     main = "Scatterplot of price against number of bathrooms",
     xlab = "number of bathrooms", ylab = "price",
     col="red",
     pch = 19, frame = FALSE)
```




```
housing <- read.csv(file="housing_v2.csv", header=TRUE, sep=",")

# converting appropriate variables to factors
housing <- within(housing, {
   view <- factor(view)
   backyard <- factor(backyard)
})

plot(housing$view, housing$price,
     main = "Scatterplot of price against view",
     xlab = "view", ylab = "price",
     col="red",
     pch = 19, frame = FALSE)
```




```r
# Subsetting data to only include the variables that are needed
myvars <- c("price", "sqft_living", "age")
housing_subset <- housing[myvars]

# Create the model
model1 <- lm(price ~ sqft_living + age, data=housing_subset)
summary(model1)

# fitted values for model 1
fitted_values <- fitted.values(model1)

# residuals for model 1
residuals <- residuals(model1)
```

```
Call:
lm(formula = price ~ sqft_living + age, data = housing_subset)

Residuals:
    Min      1Q  Median      3Q     Max 
-427747 -110061   -8090  101450  537643 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11470.393  11013.426  -1.041    0.298    
sqft_living    215.829      4.115  52.449   <2e-16 ***
age           1439.334    107.045  13.446   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 148500 on 2689 degrees of freedom
Multiple R-squared:  0.5084,	Adjusted R-squared:  0.5081 
F-statistic:  1391 on 2 and 2689 DF,  p-value: < 2.2e-16
```

Model #2 - Complete Second Order Regression Model with Quantitative Variables

A complete second order regression model was created for price as the response variable, and school_rating and crime as predictor variables.

```
# data includes only needed variables
print("Second Order Regression Model for Model 2")
myvars <- c("price", "school_rating", "crime")
housing_subset <- housing[myvars]

# Create second order regression model
model2 <- lm(price ~ school_rating + crime + school_rating:crime + I(school_rating^2) + I(crime^2), data=housing_subset)
summary(model2)
```

```
[1] "Second Order Regression Model for Model 2"

Call:
lm(formula = price ~ school_rating + crime + school_rating:crime +
    I(school_rating^2) + I(crime^2), data = housing_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-340729  -61055   -6288   56875  427915

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.339e+05  1.032e+05   7.113 1.45e-12 ***
school_rating      -7.375e+04  2.083e+04  -3.541 0.000406 ***
crime              -3.155e+03  5.235e+02  -6.027 1.90e-09 ***
I(school_rating^2)  1.165e+04  1.109e+03  10.497  < 2e-16 ***
I(crime^2)          6.377e+00  7.265e-01   8.777  < 2e-16 ***
school_rating:crime -5.227e+01  4.853e+01  -1.077 0.281513
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 92690 on 2686 degrees of freedom
Multiple R-squared:  0.8088,	Adjusted R-squared:  0.8084
F-statistic:  2272 on 5 and 2686 DF,  p-value: < 2.2e-16
```

Nested Models F-Test

This reduced model is compared with the complete second order model (Model #2 above)

```
# data includes only needed variables
print("Nested Model for Model 2")
myvars <- c("price", "school_rating", "crime")
housing_subset <- housing[myvars]

# this is the reduced model for model 2
model2_reduced <- lm(price ~ school_rating + crime + school_rating:crime, data=housing)
summary(model2_reduced)

# The Nested Model F-test
anova(model2, model2_reduced)
```

```
[1] "Nested Model for Model 2"

Call:
lm(formula = price ~ school_rating + crime + school_rating:crime, 
    data = housing)

Residuals:
    Min      1Q  Median      3Q     Max 
-336984  -63754   -4397   58894  440377 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -410233.37   25261.25  -16.24   <2e-16 ***
school_rating        155559.97    3133.06   49.65   <2e-16 ***
crime                  2230.07     129.70   17.20   <2e-16 ***
school_rating:crime    -564.85      17.86  -31.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 94870 on 2688 degrees of freedom
Multiple R-squared:  0.7995,	Adjusted R-squared:  0.7993 
F-statistic:  3573 on 3 and 2688 DF,  p-value: < 2.2e-16
```

A anova: 2 × 6

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2686	2.307469e+13	NA	NA	NA	NA
2688	2.419501e+13	-2	-1.120319e+12	65.20513	2.22716e-28