R Notebook

Importing the dataset

```
library(readxl)
uffidata <- read_excel("C:/Users/Sagar Ghiya/Desktop/Study/Sem 4/Intro to DMML/Practicum2/uffidata.xlsx
df <- data.frame(uffidata)
```

Exploring the structure of the data

```
str(df)
```

```
## 'data.frame':    99 obs. of  12 variables:
##  $ Observation   : num  37 79 75 32 69 4 28 30 18 63 ...
##  $ Year.Sold     : num  2009 2009 2011 2011 2010 ...
##  $ Sale.Price    : num  76900 78000 79000 80000 82000 84000 84000 84000 85000 85000 ...
##  $ UFFI.IN       : num  1 1 0 0 1 1 0 0 0 1 ...
##  $ Brick.Ext     : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ X45.Yrs.      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Bsmnt.Fin_SF  : num  0 154 400 0 157 ...
##  $ Lot.Area      : num  2772 4490 5840 5040 5441 ...
##  $ Enc.Pk.Spaces : num  0 0 0 0 0 1 2 0 0 1 ...
##  $ Living.Area_SF: num  1018 536 721 513 672 ...
##  $ Central.Air   : num  0 1 1 0 0 0 0 0 1 0 ...
##  $ Pool          : num  0 0 0 0 0 0 0 0 0 0 ...
```

Part1

Checking for outliers. For this problem, outliers are those which are 3 standard deviations apart from mean. Figuring out attributes that have outliers.

```
k <- vector()
for(i in 1:ncol(df)) {
    k[i] <- sum(abs((df[,i] - mean(df[,i]))/sd(df[,i]))>3)
}
k
```

```
##  [1] 0 0 3 0 0 0 0 0 0 2 0 3
```

Yes there are outliers in the dataset.

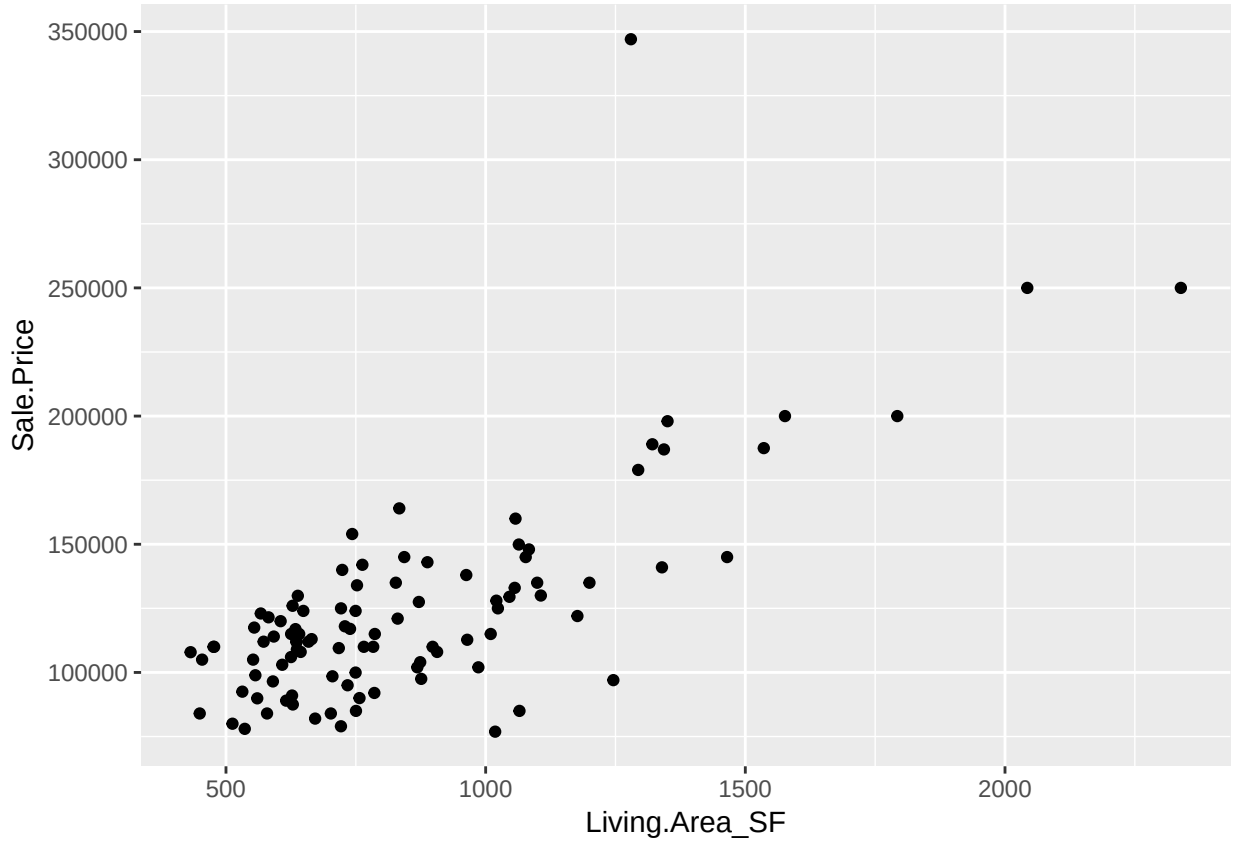
From above results, Columns 3, 10 and 12 have outliers.

So sale price, living area and pool have outliers. Now pool only has values 0 and 1. It shows outliers due to very few values of '1'. So we can ignore outliers for pool attribute.

To figure out outliers for sale price and living area, let's plot a scatter plot to see which values would actually affect our analysis.

From the scatter plot we observe that there is only 1 value of sale price and living area_SF which is very far and might ruin our analysis. So removing that particular row from dataset.

```
library(ggplot2)
ggplot(df, aes(Living.Area_SF, Sale.Price)) + geom_point()
```




```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

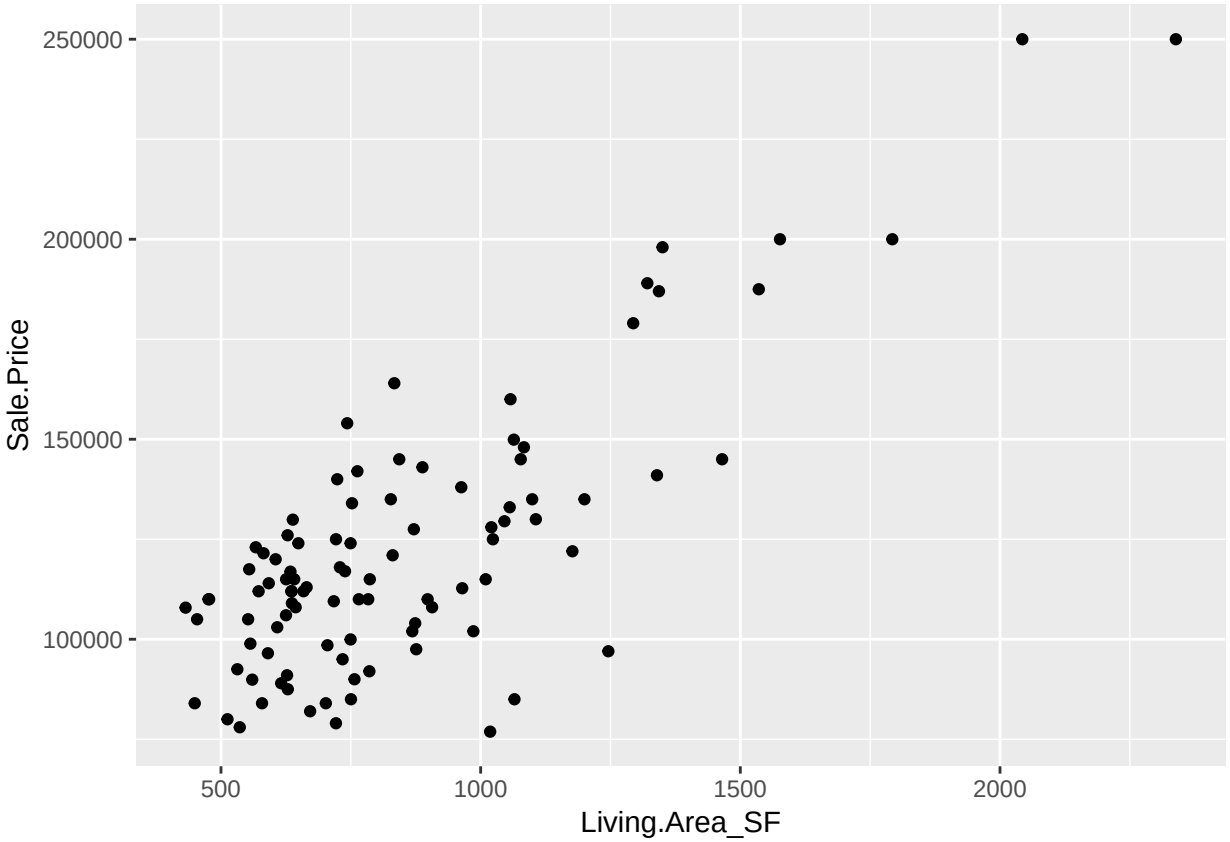
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df1 <- df %>% filter(Sale.Price<300000)
```

Exploring again. Now it can be seen that all values look decent to fit our multiple regression model.

Thus we have succesfully dealt with outliers.

```
ggplot(df1, aes(Living.Area_SF, Sale.Price)) + geom_point()
```



#Part2

pairs.panels is ideal function for this question as it serves all purposes. It gives histogram to check for normality, correlation between response variable and other features and correlation between features(collinearity).

Plotting with pairs.panels for visualization. Sale price is distributed normally and hence is suitable for parametric approaches.

As it can be seen sale price is decently correlated with some features such as living space and not so much with a few. However less correlated features will be removed when we implement backfitting.

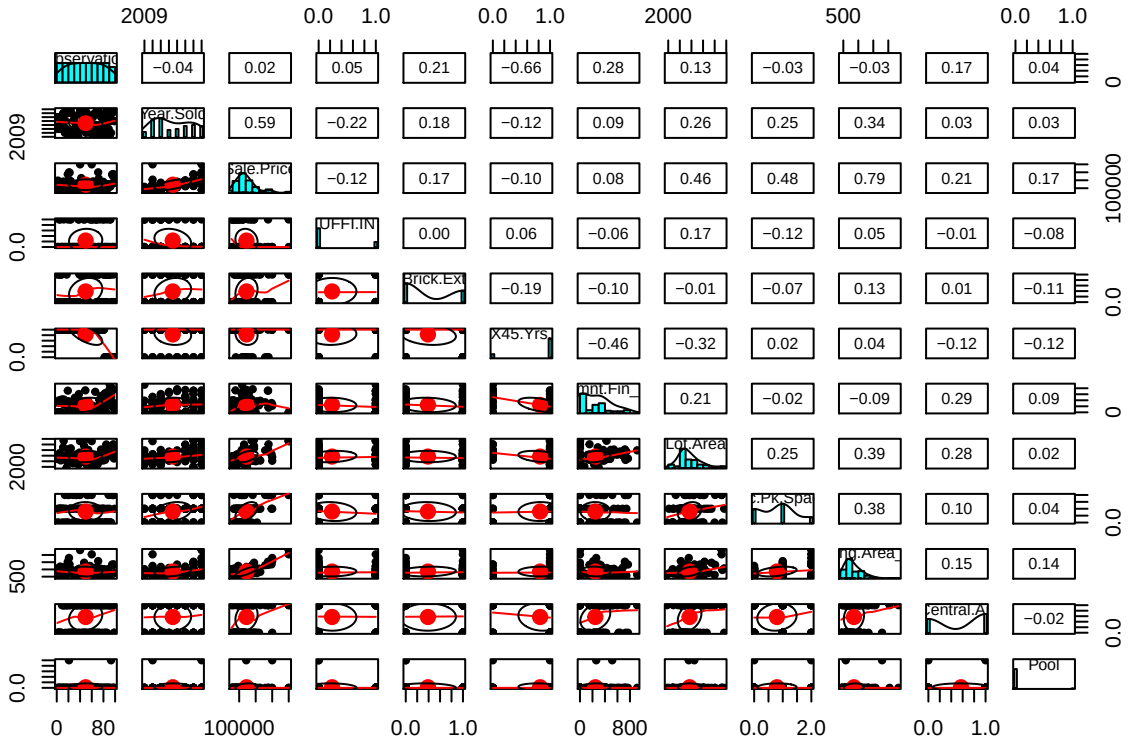
The predictor variables are not extremely correlated. There are no two variables with correlation more than 0.5 or less than -0.5. So collinearity should not be a problem here.

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
pairs.panels(df1)
```



Part3

No, the presence or absence of UFFI alone is not enough to predict sale price as the correlation between two is only -0.12. There are other features which are more correlated and infact more important than UFFI for prediction.

Part4

Implementing the multiple regression model and using backfitting to determine significant features using AIC.

```
model <- step(lm(Sale.Price ~. , data=df1), type='backward')
```

```
## Start:  AIC=1908.04
## Sale.Price ~ Observation + Year.Sold + UFFI.IN + Brick.Ext +
##     X45.Yrs. + Bsmnt.Fin_SF + Lot.Area + Enc.Pk.Spaces + Living.Area_SF +
##     Central.Air + Pool
##
##                  Df  Sum of Sq        RSS    AIC
## - Observation     1 3.7931e+06 2.1906e+10 1906.1
## - X45.Yrs.        1 8.3926e+06 2.1910e+10 1906.1
## - Central.Air     1 1.3102e+08 2.2033e+10 1906.6
## - Bsmnt.Fin_SF    1 3.5094e+08 2.2253e+10 1907.6
## - Pool            1 3.7625e+08 2.2278e+10 1907.7
```

```
## - Brick.Ext       1 3.9828e+08 2.2300e+10 1907.8
## <none>                         2.1902e+10 1908.0
## - UFFI.IN         1 6.6055e+08 2.2563e+10 1909.0
## - Lot.Area        1 9.6711e+08 2.2869e+10 1910.3
## - Enc.Pk.Spaces   1 1.8620e+09 2.3764e+10 1914.0
## - Year.Sold       1 6.5481e+09 2.8450e+10 1931.7
## - Living.Area_SF  1 2.3862e+10 4.5764e+10 1978.3
##
## Step:  AIC=1906.06
## Sale.Price ~ Year.Sold + UFFI.IN + Brick.Ext + X45.Yrs. + Bsmnt.Fin_SF +
##     Lot.Area + Enc.Pk.Spaces + Living.Area_SF + Central.Air +
##     Pool
##
##                   Df  Sum of Sq        RSS    AIC
## - X45.Yrs.         1 4.6451e+06 2.1910e+10 1904.1
## - Central.Air      1 1.4092e+08 2.2047e+10 1904.7
## - Bsmnt.Fin_SF     1 3.4982e+08 2.2256e+10 1905.6
## - Pool             1 3.7399e+08 2.2280e+10 1905.7
## - Brick.Ext        1 4.1292e+08 2.2319e+10 1905.9
## <none>                          2.1906e+10 1906.1
## - UFFI.IN          1 6.5818e+08 2.2564e+10 1907.0
## - Lot.Area         1 9.6981e+08 2.2876e+10 1908.3
## - Enc.Pk.Spaces    1 1.8748e+09 2.3781e+10 1912.1
## - Year.Sold        1 6.6216e+09 2.8527e+10 1929.9
## - Living.Area_SF   1 2.3932e+10 4.5838e+10 1976.4
##
## Step:  AIC=1904.08
## Sale.Price ~ Year.Sold + UFFI.IN + Brick.Ext + Bsmnt.Fin_SF +
##     Lot.Area + Enc.Pk.Spaces + Living.Area_SF + Central.Air +
##     Pool
##
##                   Df  Sum of Sq        RSS    AIC
## - Central.Air      1 1.4530e+08 2.2056e+10 1902.7
## - Pool             1 3.6999e+08 2.2280e+10 1903.7
## - Bsmnt.Fin_SF     1 3.8468e+08 2.2295e+10 1903.8
## - Brick.Ext        1 4.2600e+08 2.2336e+10 1904.0
## <none>                          2.1910e+10 1904.1
## - UFFI.IN          1 6.5367e+08 2.2564e+10 1905.0
## - Lot.Area         1 1.0382e+09 2.2949e+10 1906.6
## - Enc.Pk.Spaces    1 1.8813e+09 2.3792e+10 1910.2
## - Year.Sold        1 6.6287e+09 2.8539e+10 1928.0
## - Living.Area_SF   1 2.4611e+10 4.6522e+10 1975.9
##
## Step:  AIC=1902.72
## Sale.Price ~ Year.Sold + UFFI.IN + Brick.Ext + Bsmnt.Fin_SF +
##     Lot.Area + Enc.Pk.Spaces + Living.Area_SF + Pool
##
##                   Df  Sum of Sq        RSS    AIC
## - Pool             1 3.3990e+08 2.2396e+10 1902.2
## - Brick.Ext        1 4.4720e+08 2.2503e+10 1902.7
## <none>                          2.2056e+10 1902.7
## - Bsmnt.Fin_SF     1 5.6700e+08 2.2623e+10 1903.2
## - UFFI.IN          1 6.9390e+08 2.2750e+10 1903.8
## - Lot.Area         1 1.2305e+09 2.3286e+10 1906.0
```

```
## - Enc.Pk.Spaces   1 1.9206e+09 2.3976e+10 1908.9
## - Year.Sold       1 6.4934e+09 2.8549e+10 1926.0
## - Living.Area_SF  1 2.5399e+10 4.7455e+10 1975.8
##
## Step:  AIC=1902.22
## Sale.Price ~ Year.Sold + UFFI.IN + Brick.Ext + Bsmnt.Fin_SF +
##     Lot.Area + Enc.Pk.Spaces + Living.Area_SF
##
##                  Df  Sum of Sq        RSS    AIC
## - Brick.Ext       1 3.5785e+08 2.2754e+10 1901.8
## <none>                         2.2396e+10 1902.2
## - Bsmnt.Fin_SF    1 6.6118e+08 2.3057e+10 1903.1
## - UFFI.IN         1 7.7988e+08 2.3176e+10 1903.6
## - Lot.Area        1 1.1750e+09 2.3571e+10 1905.2
## - Enc.Pk.Spaces   1 1.8696e+09 2.4265e+10 1908.1
## - Year.Sold       1 6.4408e+09 2.8837e+10 1925.0
## - Living.Area_SF  1 2.7355e+10 4.9750e+10 1978.4
##
## Step:  AIC=1901.78
## Sale.Price ~ Year.Sold + UFFI.IN + Bsmnt.Fin_SF + Lot.Area +
##     Enc.Pk.Spaces + Living.Area_SF
##
##                  Df  Sum of Sq        RSS    AIC
## <none>                         2.2754e+10 1901.8
## - Bsmnt.Fin_SF    1 5.7941e+08 2.3333e+10 1902.2
## - UFFI.IN         1 7.5102e+08 2.3505e+10 1903.0
## - Lot.Area        1 1.1137e+09 2.3867e+10 1904.5
## - Enc.Pk.Spaces   1 1.6785e+09 2.4432e+10 1906.8
## - Year.Sold       1 7.2578e+09 3.0011e+10 1926.9
## - Living.Area_SF  1 2.8423e+10 5.1176e+10 1979.2
```

Thus we get following as our optimal model

```
summary(model)
```

```
##
## Call:
## lm(formula = Sale.Price ~ Year.Sold + UFFI.IN + Bsmnt.Fin_SF +
##     Lot.Area + Enc.Pk.Spaces + Living.Area_SF, data = df1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33187  -6618    211   7550  55568
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -8.680e+06  1.619e+06  -5.362 6.20e-07 ***
## Year.Sold       4.339e+03  8.054e+02   5.388 5.56e-07 ***
## UFFI.IN        -7.018e+03  4.050e+03  -1.733   0.0865 .
## Bsmnt.Fin_SF    1.072e+01  7.044e+00   1.522   0.1314
## Lot.Area        1.947e+00  9.225e-01   2.111   0.0376 *
## Enc.Pk.Spaces   6.707e+03  2.589e+03   2.591   0.0111 *
## Living.Area_SF  6.070e+01  5.693e+00  10.662  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 15810 on 91 degrees of freedom
## Multiple R-squared:  0.7921, Adjusted R-squared:  0.7784
## F-statistic: 57.79 on 6 and 91 DF,  p-value: < 2.2e-16
```

Yes UFFI is a significant predictor variable when taken with full set of variables. This can be observed from above results.

Part 5

Multiple regression model is built above. The model is

Sale Price = -8.680e+06 + 4.339e+03 *Year.Sold* - *7.018e+03*UFFI.IN + 1.072e+01*Bsmnt.Fin_SF* + *1.947e+00*Lot.Area + 6.707e+03*Enc.Pk.Spaces* + *6.070e+01*Living.Area_SF

Adjusted R squared = 0.7784. P values for principal components can be observed from the above summary.

Following gives the RMSE of our model.

```
RMSE <- function(residuals) {
  sqrt(mean(residuals^2))
}
```

```
RMSE(model$residuals)
```

```
## [1] 15237.42
```

Thus RMSE of the model is 15237.42

Part6

Regression equation is:

Sale.Price = -8.680e+06 + 4.339e+03 *Year.Sold* - *7.018e+03*UFFI.IN + 1.072e+01*Bsmnt.Fin_SF* + *1.947*Lot.Area + 6.707e+03*Enc.Pk.Spaces* + *6.070e+01*Living.Area_SF

On average, 1 unit of UFFI will decrease the price of property by $7018

Part7

From the model it can be observed that year sold is an important criteria to predict sale price. However it has not been given. So putting the value of year as 2016 which is most recent.

```
table(df1$Year.Sold)
```

```
##
## 2009 2010 2011 2012 2013 2014 2015 2016
##    6   18   20    8    9   12   13   12
```

Without UFFI

```
newdata1 <- data.frame(2016,0,1,1,0.000,5000,2,1700,1,0)
colnames(newdata1) <- c('Year.Sold','UFFI.IN', 'Brick.ext', 'X45.Yrs', 'Bsmnt.Fin_SF', 'Lot.Area','Enc.
```

Making predictions

```
ans1 <- predict(model,newdata1)
ans1
```

```
##        1
## 194597.2
```

Calculating 95 % confidence intervals.

```
low_bound <- ans1 - 1.96*15810
high_bound <- ans1 + 1.96*15810
low_bound
```

```
##        1
## 163609.6
```

```
high_bound
```

```
##        1
## 225584.8
```

WithUFFI

```
newdata2 <- data.frame(2016,1,1,1,0.000,5000,2,1700,1,0)
colnames(newdata2) <- c('Year.Sold','UFFI.IN', 'Brick.ext', 'X45.Yrs', 'Bsmnt.Fin_SF', 'Lot.Area','Enc.
```

Prediction of sale price with UFFI

```
ans2 <- predict(model,newdata2)
ans2
```

```
##        1
## 187578.8
```

Calculating 95% confidence intervals.

```
l_bound <- ans2 - 1.96*15810
h_bound <- ans2 + 1.96*15810
l_bound
```

```
##        1
## 156591.2
```

```
h_bound
```

```
##        1
## 218566.4
```

Part8

Client overpayed: 215000 - 187578.8 = $27,421.2 So compensation of $27,421.2 is justified.