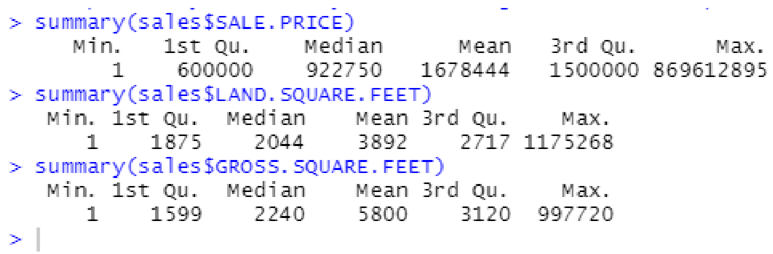
ITWS - 6600 Data Analytics

Assignment 4: Pattern, Trends, Relations

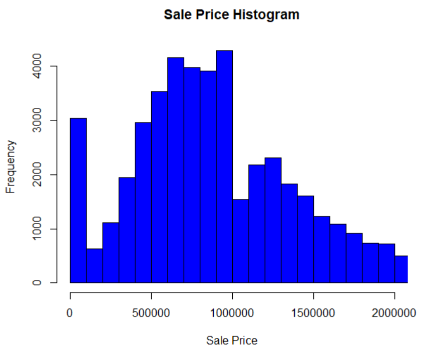
Submitted by: Shrey Jain (RIN: 662046332)

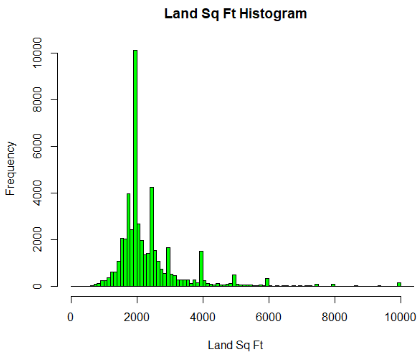
Q.1

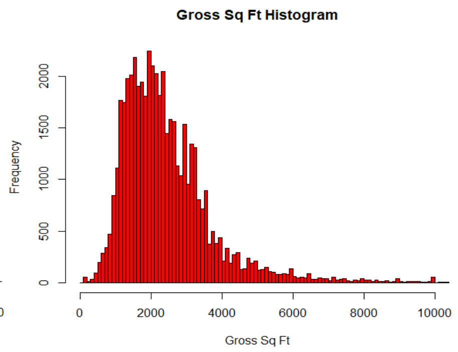
(a) All of my analysis for this assignment was done on the Brooklyn borough. The three columns that I chose for my analysis are sale price, land square feet and gross square feet. Firstly, the columns land and gross square feet contained comma separated numeric values, hence it was required to remove those commas and convert that to numeric in R. I also removed NA values from the data. After which, I used summary functions to identify the 4 Quartiles. I used histograms to see the patterns and trends in the selected columns.

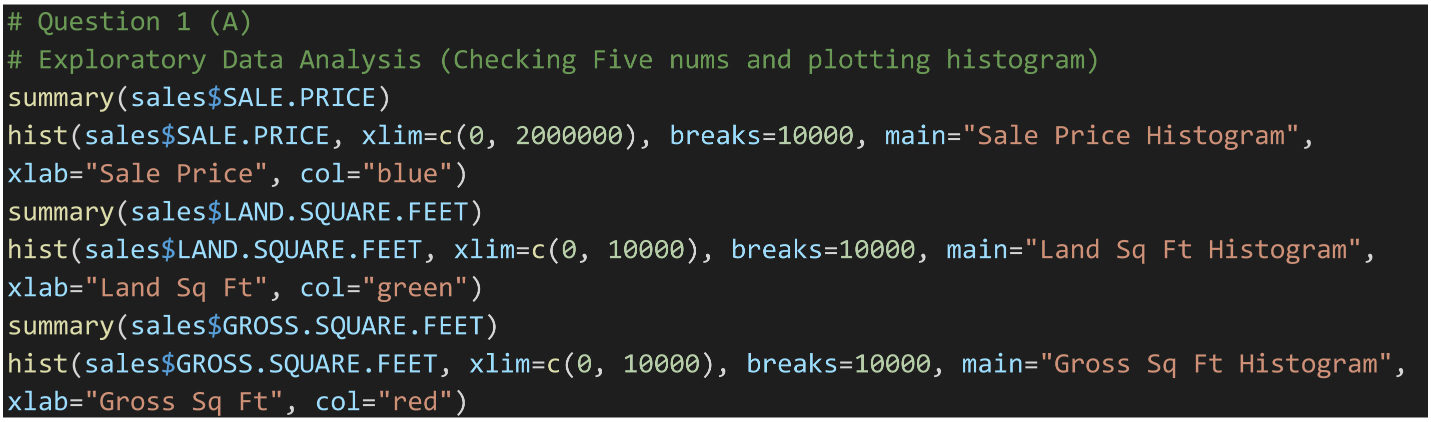


```
> summary(sales$SALE.PRICE)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
        1    600000    922750   1678444   1500000 869612895 
> summary(sales$LAND.SQUARE.FEET)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      1    1875    2044    3892    2717 1175268 
> summary(sales$GROSS.SQUARE.FEET)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      1    1599    2240    5800    3120  997720 
> 
```



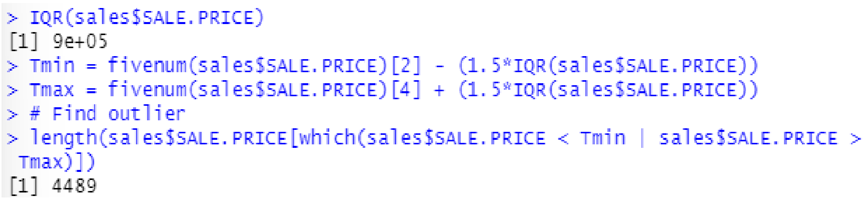




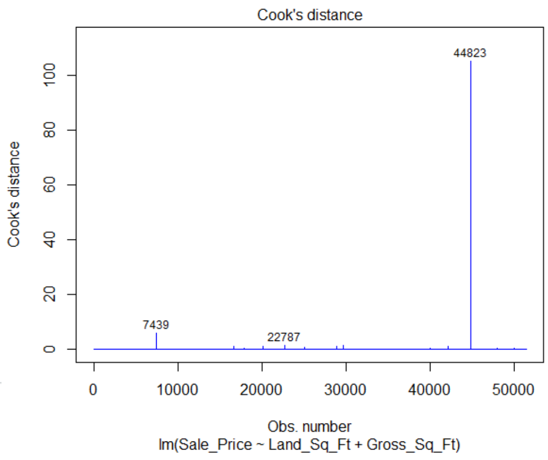


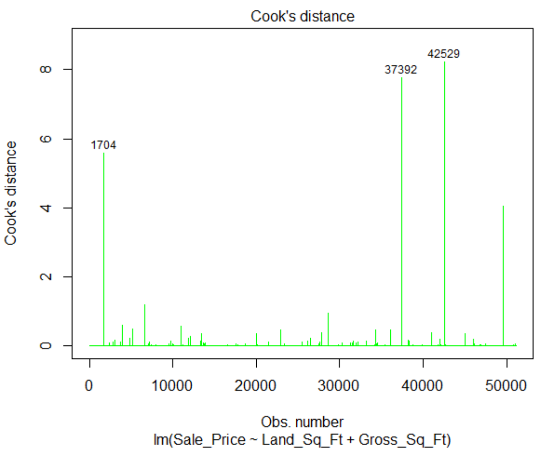
```
# Question 1 (A)
# Exploratory Data Analysis (Checking Five nums and plotting histogram)
summary(sales$SALE.PRICE)
hist(sales$SALE.PRICE, xlim=c(0, 2000000), breaks=10000, main="Sale Price Histogram",
xlab="Sale Price", col="blue")
summary(sales$LAND.SQUARE.FEET)
hist(sales$LAND.SQUARE.FEET, xlim=c(0, 10000), breaks=10000, main="Land Sq Ft Histogram",
xlab="Land Sq Ft", col="green")
summary(sales$GROSS.SQUARE.FEET)
hist(sales$GROSS.SQUARE.FEET, xlim=c(0, 10000), breaks=10000, main="Gross Sq Ft Histogram",
xlab="Gross Sq Ft", col="red")
```

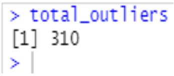
(b) As mentioned in the class, to identify outliers, I used both the IQR method as well as the Cook's distance method. Using the IQR method, the number of outliers found for sale.price column were 4489 out of 51460. Firstly, I calculated the IQR which came out to be 900000. The I used the lower and upper bound formulas to calculate them. Finally, I found the number of data points that lie outside these bounds. The same steps were repeated for other two columns which are land.square.feet and gross.square.feet
For the second approach, using the cook's distance method, I followed the code snippets from the lecture slides. The total points came out to be 310 of the model.



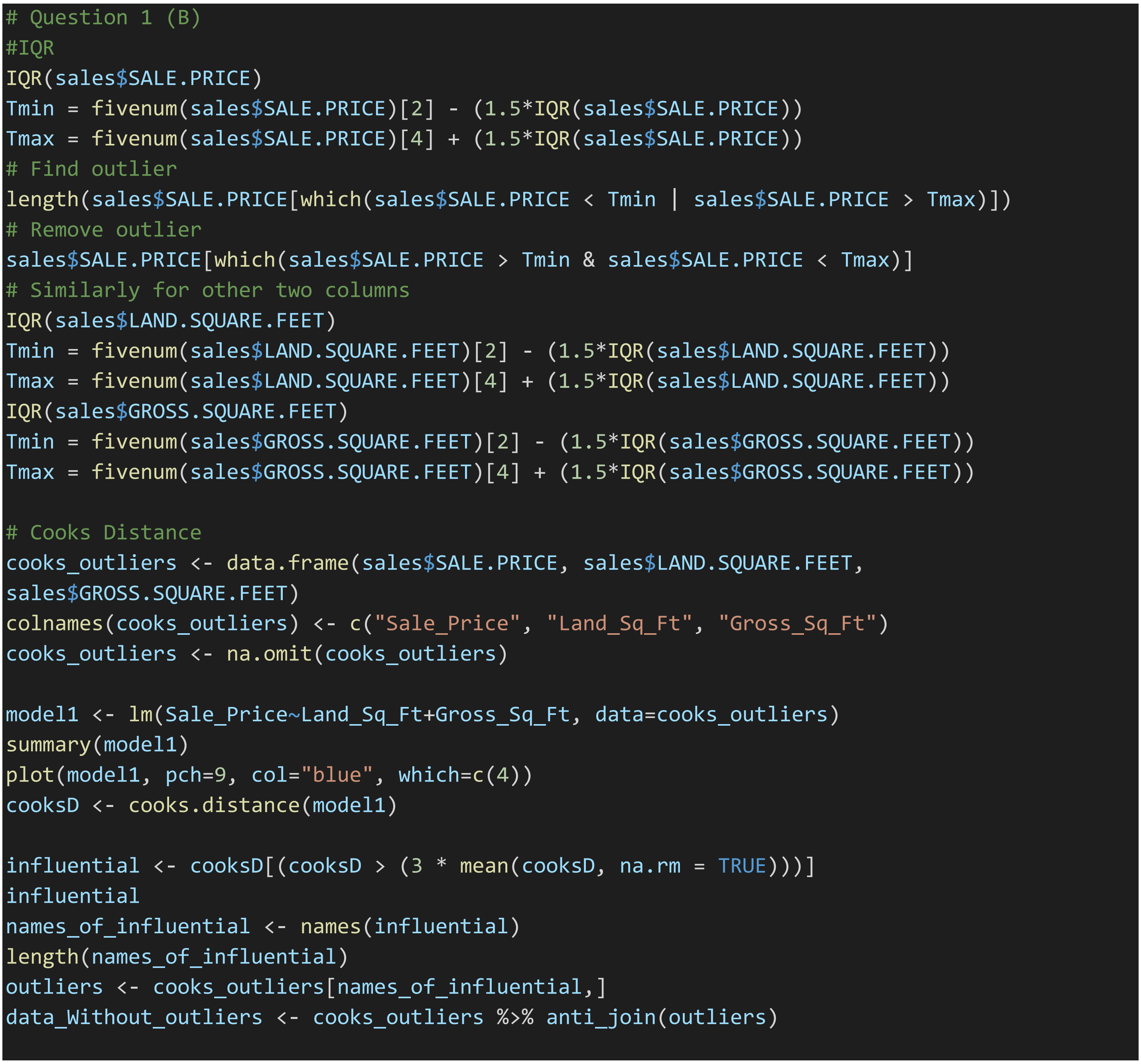
```
> IQR(sales$SALE.PRICE)
[1] 9e+05
> Tmin = fivenum(sales$SALE.PRICE)[2] - (1.5*IQR(sales$SALE.PRICE))
> Tmax = fivenum(sales$SALE.PRICE)[4] + (1.5*IQR(sales$SALE.PRICE))
> # Find outlier
> length(sales$SALE.PRICE[which(sales$SALE.PRICE < Tmin | sales$SALE.PRICE >
 Tmax)])
[1] 4489
```







```
> total_outliers
[1] 310
> |
```



```
# Question 1 (B)
#IQR
IQR(sales$SALE.PRICE)
Tmin = fivenum(sales$SALE.PRICE)[2] - (1.5*IQR(sales$SALE.PRICE))
Tmax = fivenum(sales$SALE.PRICE)[4] + (1.5*IQR(sales$SALE.PRICE))
# Find outlier
length(sales$SALE.PRICE[which(sales$SALE.PRICE < Tmin | sales$SALE.PRICE > Tmax)])
# Remove outlier
sales$SALE.PRICE[which(sales$SALE.PRICE > Tmin & sales$SALE.PRICE < Tmax)]
# Similarly for other two columns
IQR(sales$LAND.SQUARE.FEET)
Tmin = fivenum(sales$LAND.SQUARE.FEET)[2] - (1.5*IQR(sales$LAND.SQUARE.FEET))
Tmax = fivenum(sales$LAND.SQUARE.FEET)[4] + (1.5*IQR(sales$LAND.SQUARE.FEET))
IQR(sales$GROSS.SQUARE.FEET)
Tmin = fivenum(sales$GROSS.SQUARE.FEET)[2] - (1.5*IQR(sales$GROSS.SQUARE.FEET))
Tmax = fivenum(sales$GROSS.SQUARE.FEET)[4] + (1.5*IQR(sales$GROSS.SQUARE.FEET))

# Cooks Distance
cooks_outliers <- data.frame(sales$SALE.PRICE, sales$LAND.SQUARE.FEET,
sales$GROSS.SQUARE.FEET)
colnames(cooks_outliers) <- c("Sale_Price", "Land_Sq_Ft", "Gross_Sq_Ft")
cooks_outliers <- na.omit(cooks_outliers)

model1 <- lm(Sale_Price~Land_Sq_Ft+Gross_Sq_Ft, data=cooks_outliers)
summary(model1)
plot(model1, pch=9, col="blue", which=c(4))
cooksD <- cooks.distance(model1)

influential <- cooksD[(cooksD > (3 * mean(cooksD, na.rm = TRUE)))]
influential
names_of_influential <- names(influential)
length(names_of_influential)
outliers <- cooks_outliers[names_of_influential,]
data_Without_outliers <- cooks_outliers %>% anti_join(outliers)
```

```
model2 <- lm(Sale_Price~Land_Sq_Ft+Gross_Sq_Ft, data=data_Without_outliers)
summary(model1)
plot(model2, pch=9, col="green", which=c(4))
cooksD2 <- cooks.distance(model2)

total_outliers <- length(cooksD) - length(cooksD2)
total_outliers
```

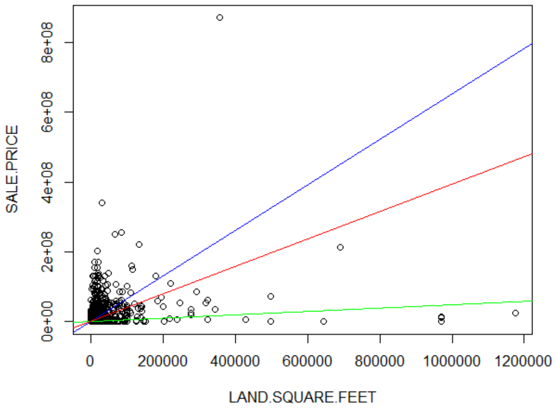
(c) For this question, I performed multivariate regression (linear, quadratic, and cubic) to predict sales price using gross square feet and land square feet. This was done in three parts, each time a different (random) sample was used using the sample function in R.

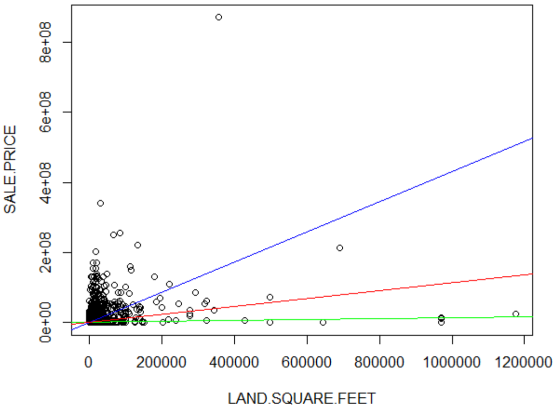
Comparison:

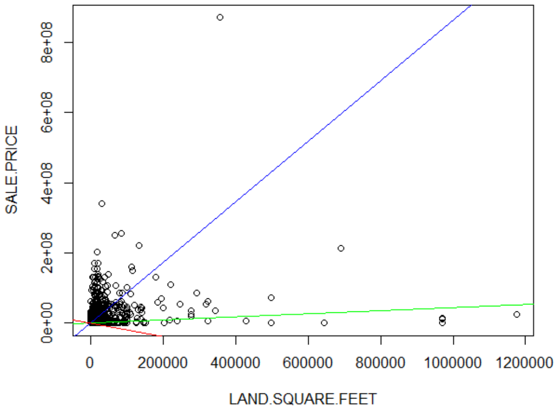
Linear (Green): Almost similar for all three rounds. While the second sample was closer towards the outliers, overall, all three samples performed the same.

Quadradic (Blue): The slope of the predicted line changed especially in sample 2. This means that more points where the sale price was low, was part of the sample 2.

Cubic (Red): First and second sample are consistent with deviation like how linear and quadratic had. However, the big thing to note is the third sample which has a drastic change in the slope. This could potentially.



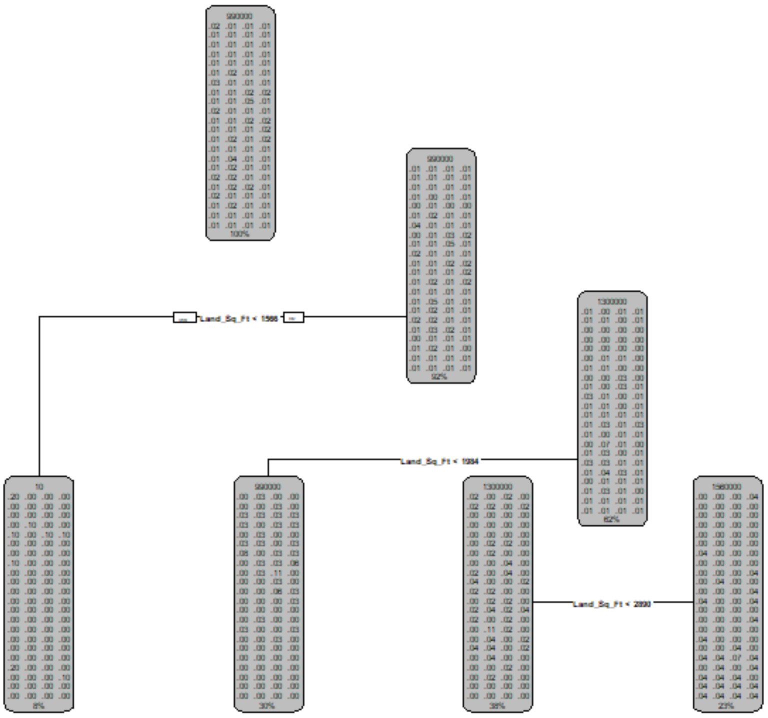


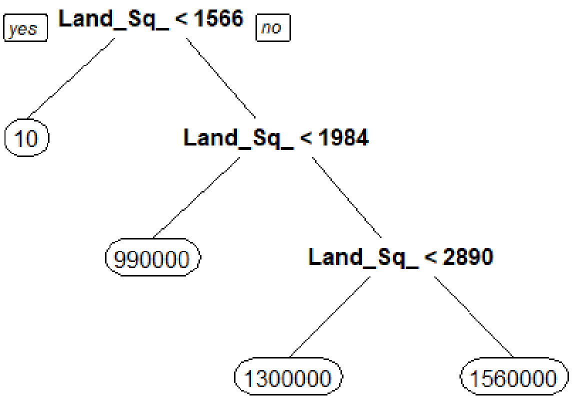


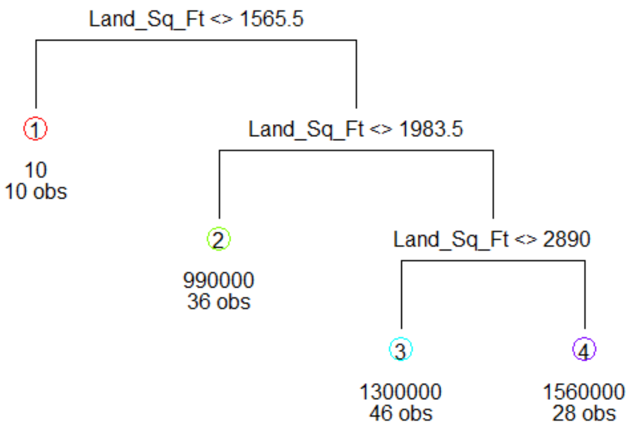
```
# Question 1 (C) Multivariate Regression
#First sample
training_dataset <- sample(dim(sales)[1], 1800)
model3_linear <- lm(SALE.PRICE~GROSS.SQUARE.FEET+LAND.SQUARE.FEET, data=sales, subset =
training_dataset)
summary(model3_linear)
plot(SALE.PRICE~GROSS.SQUARE.FEET+LAND.SQUARE.FEET, data=sales)
abline(model3_linear, col="green")

model3_quad <- lm(SALE.PRICE~poly(GROSS.SQUARE.FEET,2,raw=TRUE) +
poly(LAND.SQUARE.FEET,2,raw=TRUE), data = sales, subset = training_dataset)
```

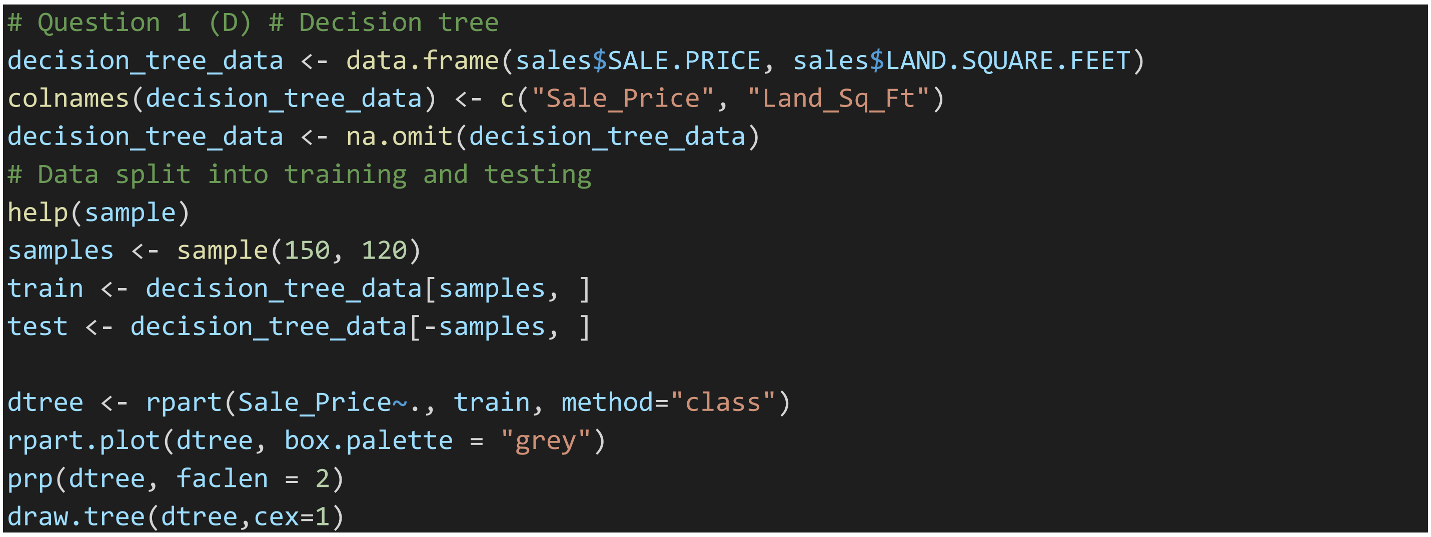
```
summary(model3_quad)
abline(model3_quad, col="blue")

model3_cube <- lm(SALE.PRICE~poly(GROSS.SQUARE.FEET,3,raw=TRUE) +
poly(LAND.SQUARE.FEET,3,raw=TRUE), data = sales, subset = training_dataset)
summary(model3_cube)
abline(model3_cube, col="red")

# Second sample
training_dataset2 <- sample(dim(sales)[1], 1800)
model4_linear <- lm(SALE.PRICE~GROSS.SQUARE.FEET+LAND.SQUARE.FEET, data=sales, subset =
training_dataset2)
summary(model4_linear)
plot(SALE.PRICE~GROSS.SQUARE.FEET+LAND.SQUARE.FEET, data=sales)
abline(model4_linear, col="green")

model4_quad <- lm(SALE.PRICE~poly(GROSS.SQUARE.FEET,2,raw=TRUE) +
poly(LAND.SQUARE.FEET,2,raw=TRUE), data = sales, subset = training_dataset2)
summary(model4_quad)
abline(model4_quad, col="blue")

model4_cube <- lm(SALE.PRICE~poly(GROSS.SQUARE.FEET,3,raw=TRUE) +
poly(LAND.SQUARE.FEET,3,raw=TRUE), data = sales, subset = training_dataset2)
summary(model4_cube)
abline(model4_cube, col="red")

# Third sample
training_dataset3 <- sample(dim(sales)[1], 1800)
model5_linear <- lm(SALE.PRICE~GROSS.SQUARE.FEET+LAND.SQUARE.FEET, data=sales, subset =
training_dataset3)
summary(model5_linear)
plot(SALE.PRICE~GROSS.SQUARE.FEET+LAND.SQUARE.FEET, data=sales)
abline(model5_linear, col="green")

model5_quad <- lm(SALE.PRICE~poly(GROSS.SQUARE.FEET,2,raw=TRUE) +
poly(LAND.SQUARE.FEET,2,raw=TRUE), data = sales, subset = training_dataset3)
summary(model5_quad)
abline(model5_quad, col="blue")

model5_cube <- lm(SALE.PRICE~poly(GROSS.SQUARE.FEET,3,raw=TRUE) +
poly(LAND.SQUARE.FEET,3,raw=TRUE), data = sales, subset = training_dataset3)
summary(model5_cube)
abline(model5_cube, col="red")
```

(d) I used decision tree for my chosen data. I was confident that decision tree might not be an ideal model for the chosen data and hence to validate my hypothesis I used it. The significance of having decision tree is especially for classification problems. While it is easier to understand and can handle multi-variable problems, it cannot supplant regression techniques. The accuracy of the decision tree came out to be extremely low (less than 1%) which confirms my initial hypothesis that decision tree would not be a good fit for my chosen data.





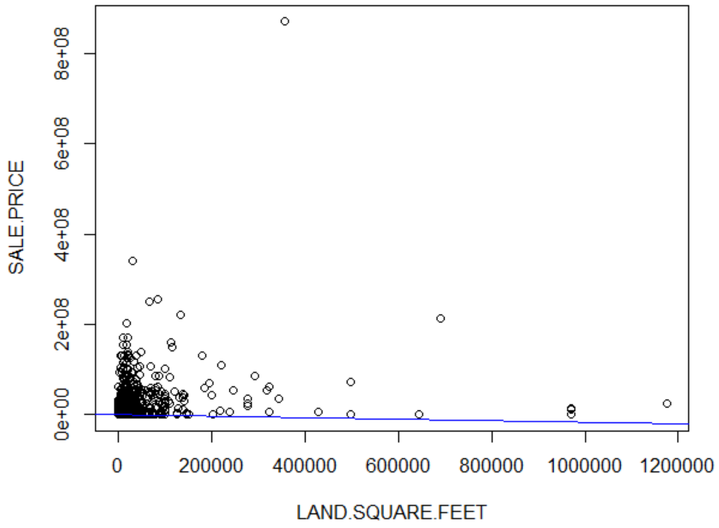




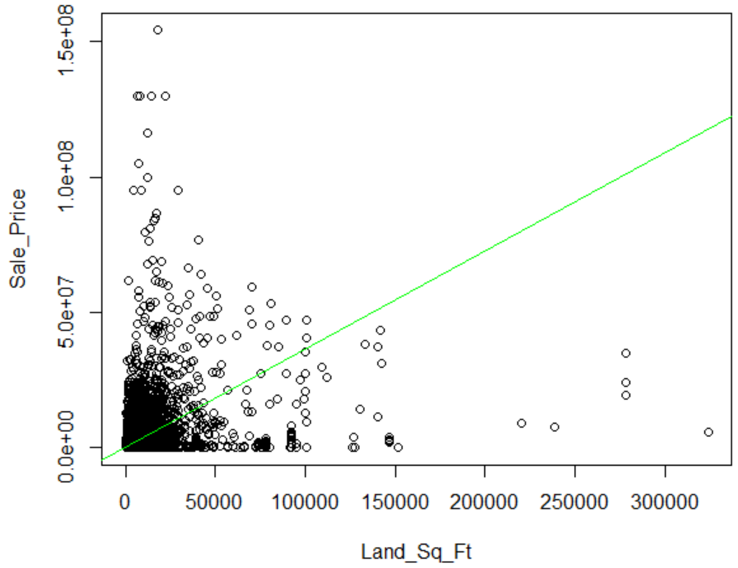
```
# Question 1 (D) # Decision tree
decision_tree_data <- data.frame(sales$SALE.PRICE, sales$LAND.SQUARE.FEET)
colnames(decision_tree_data) <- c("Sale_Price", "Land_Sq_Ft")
decision_tree_data <- na.omit(decision_tree_data)
# Data split into training and testing
help(sample)
samples <- sample(150, 120)
train <- decision_tree_data[samples, ]
test <- decision_tree_data[-samples, ]

dtree <- rpart(Sale_Price~., train, method="class")
rpart.plot(dtree, box.palette = "grey")
prp(dtree, faclen = 2)
draw.tree(dtree,cex=1)
```

Q.2

(a) On applying quadratic regression to the entire sales dataset using gross and land square feet as the determinants, we get the following plot:

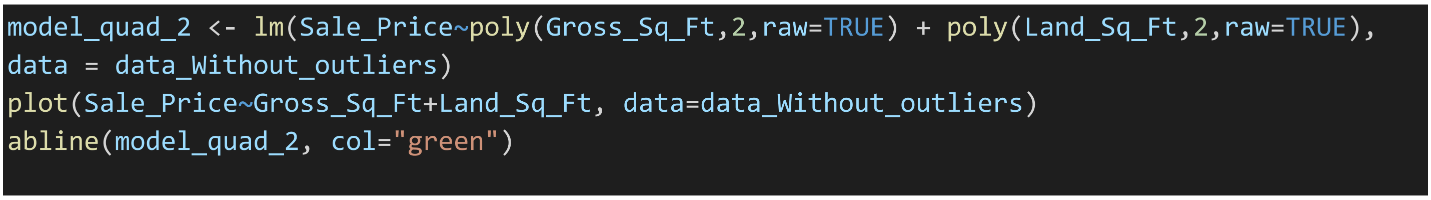


However, on using the **no outlier data** and using the same quadratic regression, we get the following plot:



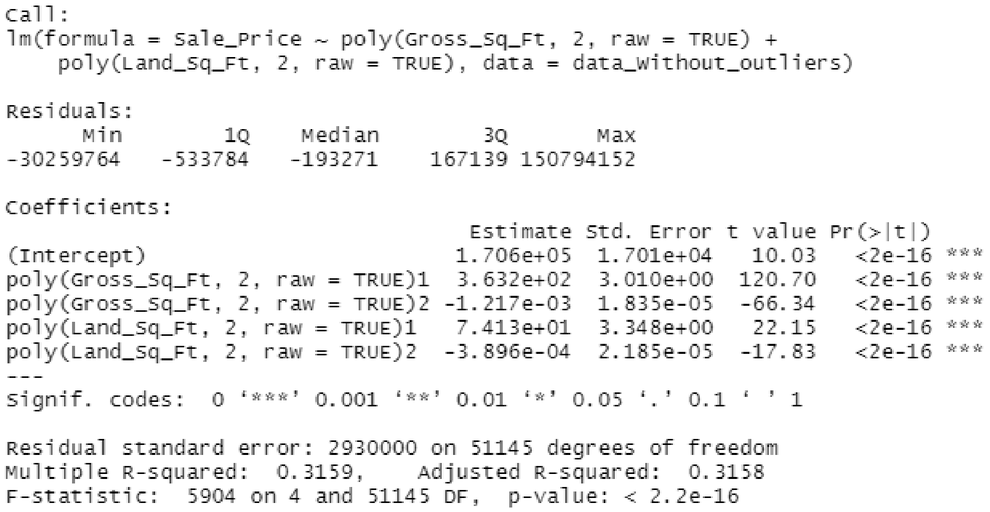
It can be seen that on fitting the model without outliers, we get better results. Also, the accuracy of second model is more than the first one. Multivariate regression does seem to predict values better, however, the data is very dense towards the initial area. The code snippet for generating the above plots is as follows:

```
# Question 2 (A)
model_quad <- lm(SALE.PRICE~poly(GROSS.SQUARE.FEET,2,raw=TRUE) +
poly(LAND.SQUARE.FEET,2,raw=TRUE), data = sales)
plot(SALE.PRICE~GROSS.SQUARE.FEET+LAND.SQUARE.FEET, data=sales)
abline(model_quad, col="blue")
```



```
model_quad_2 <- lm(Sale_Price~poly(Gross_Sq_Ft,2,raw=TRUE) + poly(Land_Sq_Ft,2,raw=TRUE),
data = data_Without_outliers)
plot(Sale_Price~Gross_Sq_Ft+Land_Sq_Ft, data=data_Without_outliers)
abline(model_quad_2, col="green")
```

(b) There are many outliers in the data which impact the accuracy of our models. On looking at the summary of our final model which was built without outliers, we get:



```
Call:
lm(formula = Sale_Price ~ poly(Gross_Sq_Ft, 2, raw = TRUE) + 
    poly(Land_Sq_Ft, 2, raw = TRUE), data = data_Without_outliers)

Residuals:
      Min        1Q    Median        3Q       Max 
-30259764   -533784   -193271    167139 150794152 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       1.706e+05  1.701e+04   10.03   <2e-16 ***
poly(Gross_Sq_Ft, 2, raw = TRUE)1  3.632e+02  3.010e+00  120.70   <2e-16 ***
poly(Gross_Sq_Ft, 2, raw = TRUE)2 -1.217e-03  1.835e-05  -66.34   <2e-16 ***
poly(Land_Sq_Ft, 2, raw = TRUE)1   7.413e+01  3.348e+00   22.15   <2e-16 ***
poly(Land_Sq_Ft, 2, raw = TRUE)2  -3.896e-04  2.185e-05  -17.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2930000 on 51145 degrees of freedom
Multiple R-squared:  0.3159,	Adjusted R-squared:  0.3158 
F-statistic:  5904 on 4 and 51145 DF,  p-value: < 2.2e-16
```

The above model summary describes all the factors/coefficients and intercepts. Looking at the p-value **2.28e^-16** is very low (almost close to 0). The R-squared error is also not that high – around **0.32**.

(c) On looking at the entire dataset, land and gross square feet attributes are not enough to perform accurate prediction. While the corelation between sales price and square feet is positive that means with increasing gross or land square feet, the sales price does go up (which also was expected, right?). Looking at other columns in the dataset, a co-relation matrix can give a better idea about which attributes play important role in determining the sales price. Also, we would need to have some negative co-related attributes as well to build a well-accurate regression model.

Q.3 The following conclusion can be drawn from my study of the NYC sales dataset.

- The summary function and histograms were helpful, especially for identifying patterns or trends and to get the general idea about the data I am working on.
- Removal of NA rows might not be the correct way for cleaning the data, a better approach could be to fill those missing values using various techniques.
- The chosen columns which are land and gross square feet might not be the sufficient attributes to predict the sales price accurately.
- Decision tree is not a good model choice to predict sales price. That is because decision trees are best suited to solve classification problems, and in our case, sales price was a numeric field for which regression is a better choice.
- Coming to regression, on performing multi variate regression, it does help cover most of the data points. Quadratic and Linear seem to perform comparatively equivalent. However, cubic is uncertain. This could potentially be just the weird sample I got in the first question, however, I tried to repeat those steps a couple of times and I did end up getting the cubic model perform inaccurately many times.