Gwyneth Yuen
ITWS 4600 - Data Analytics
October 4, 2022

Assignment 4

1. **For any one of the Brooklyn, Manhattan, Queens sales datasets, perform the following:**
    a. **Describe the type of patterns or trends you might look for and how you plan to model them. Describe any exploratory data analysis you performed. Include plots and other descriptions. Min. 5 sentences (1%)**

The sales dataset that I picked was the rolling sales in Manhattan data. The patterns or trends I could look for are how other factors affect the sale price. The one I am choosing to explore more about is how both gross square feet and year built affect the sale price. Other factors I could also compare are zip code and tax class. By looking at gross square feet, I can determine if size affects the sale price. And by looking at the year built, I can determine whether newer properties sell for a higher price. Below is the five value summaries for the sale price:

Summaries of sale price, gross square feet, and year built

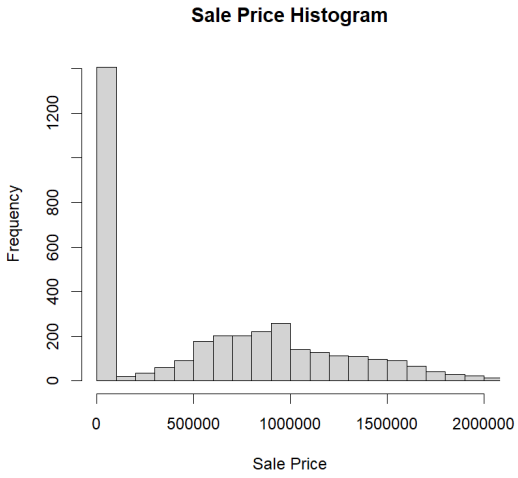
Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sale Price	0	0	730000	3267830	1318975	978090439
Gross Sq. Ft.	25	630	788	17022	3638	8942176
Year Built	1800	1915	1961	1957	1989	2019

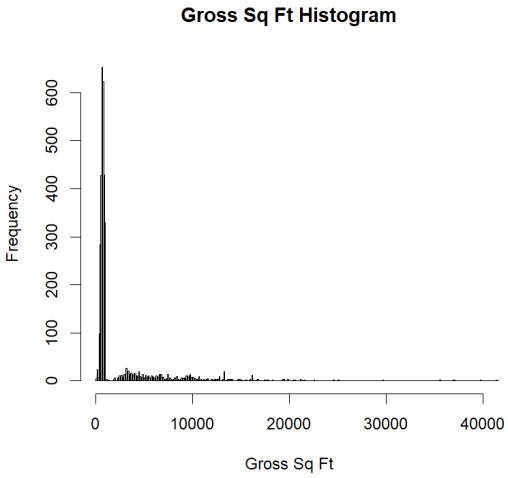
Below is the code for the above table:

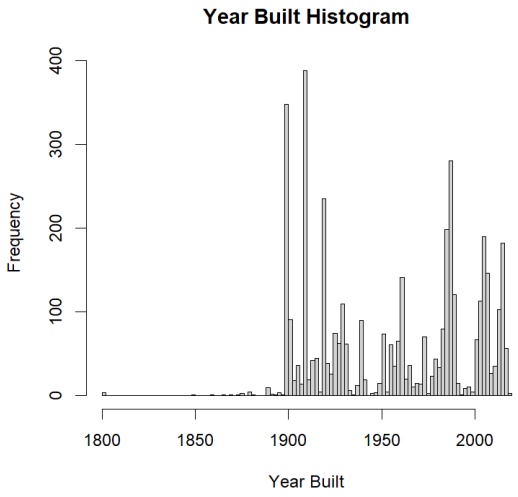
```
summary(nyc$SALE.PRICE)
summary(nyc$GROSS.SQUARE.FEET)
summary(nyc$YEAR.BUILT)
```

Histograms

```
hist(nyc$SALE.PRICE, xlim=c(0, 2000000), breaks=10000, main="Sale Price Histogram",
xlab="Sale Price")
hist(nyc$GROSS.SQUARE.FEET, xlim=c(0, 40000), breaks=100000, main="Gross Sq Ft
Histogram", xlab="Gross Sq Ft")
hist(nyc$YEAR.BUILT, breaks=100, main="Year Built Histogram", xlab="Year Built")
```







By looking at the histograms, we are able to see where the values are most populated, depending on the bin size. For example, most of the sales prices fall under $500,000, while square feet is under 10,000. Most of the properties were built after 1900.

b. Identify the outlier values in the data for Sales Price or on a variable you choose, explain why you consider those data points to be outliers? Use the Cook's Distance or IQR (InterQuartile Range) to identify the outlier points (1%)

The outlier values are as follows:

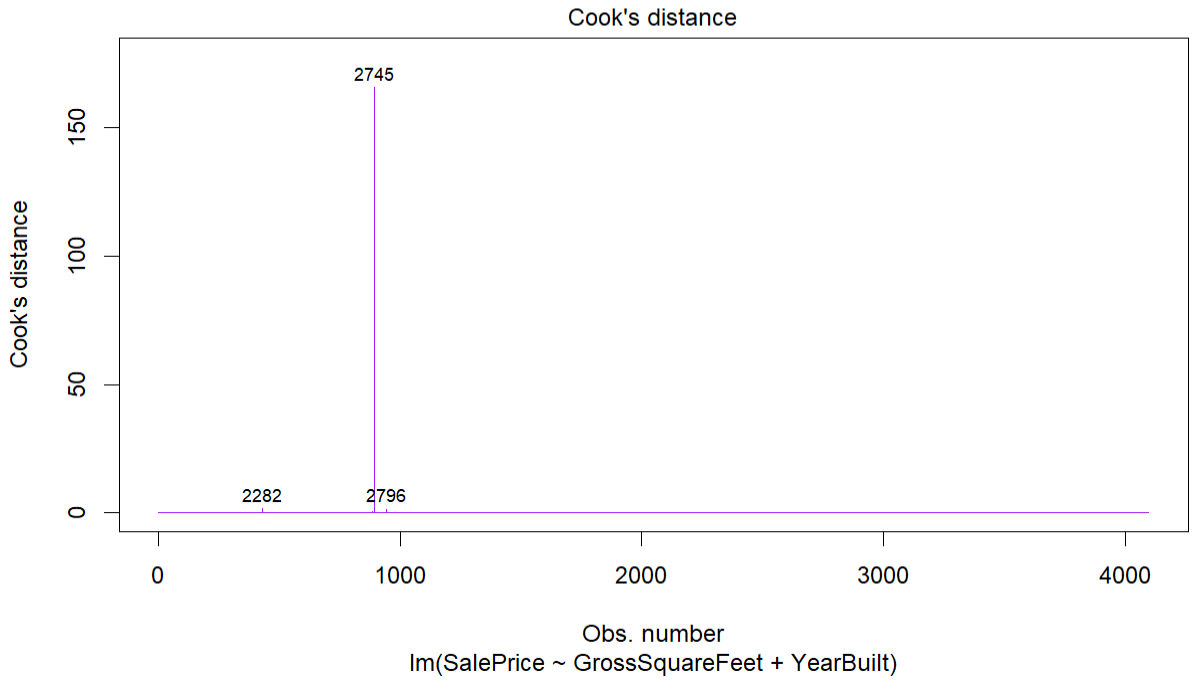
2242	2282	2739	2745	2763	2796

0.1535893	1.6925458	0.7108658	165.4275974	0.2680849	1.5285123

These values were calculated by creating a linear regression model with Sales Price as the dependent variable and Gross Square Feet and Year Built as the independent variables. A new dataframe was created for just these three variables. Then, Cook's distance was performed on this model. The outliers are the properties that were three times the mean, which comes out to six total in this instance. Below is the code for creating the model, calculating Cook's distance, and then determining the outliers 3x the mean of Cook's distance.

```
lin_reg <- lm(SalePrice~GrossSquareFeet+YearBuilt, data=outliers)
cooks <- cooks.distance(lin_reg)
influential <- cooks[cooks > (3 * mean(cooks))]
```

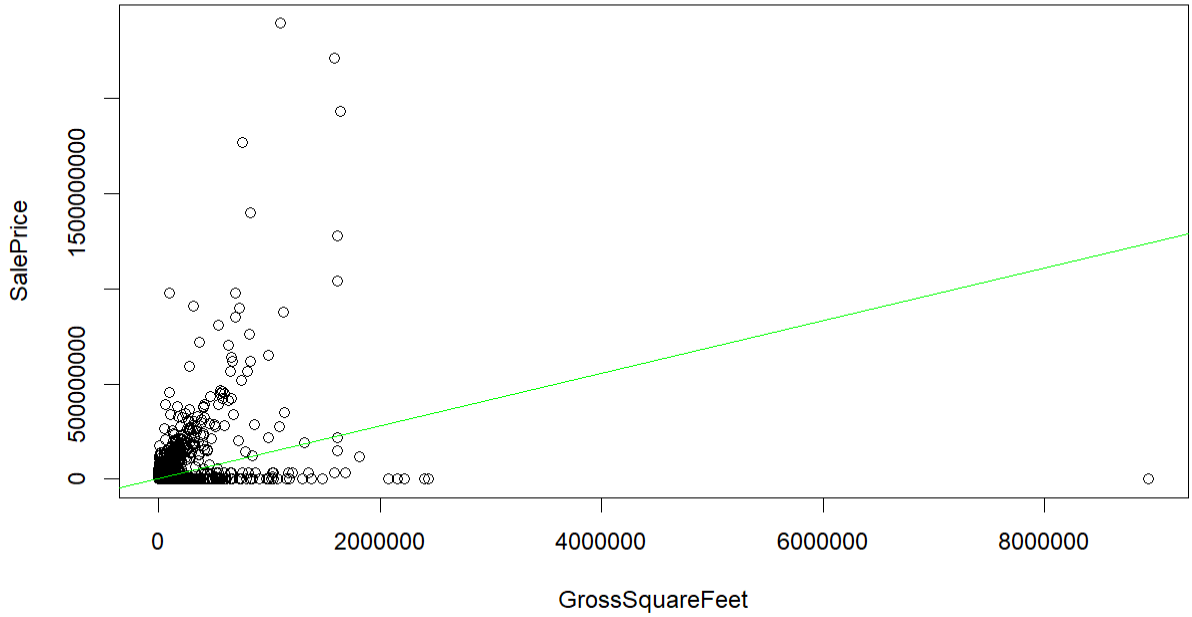
The plot for Cook's distance is:



c. Conduct Multivariate Regression on the chosen dataset to predict the Sales Price using Gross Square feet, Land Square feet. When you conduct the multivariate regression, make sure to draw at least 3 samples from the data and compare the different results you obtained. Explain the results Min. 5 sentences (1%)

When conducting multivariate regression on the dataset, the three samples chosen were: (1) sales price and gross square feet, (2) sales price and land square feet, and (3) sales price, gross square feet and land square feet, where sales prices is the dependent variable for all three samples.

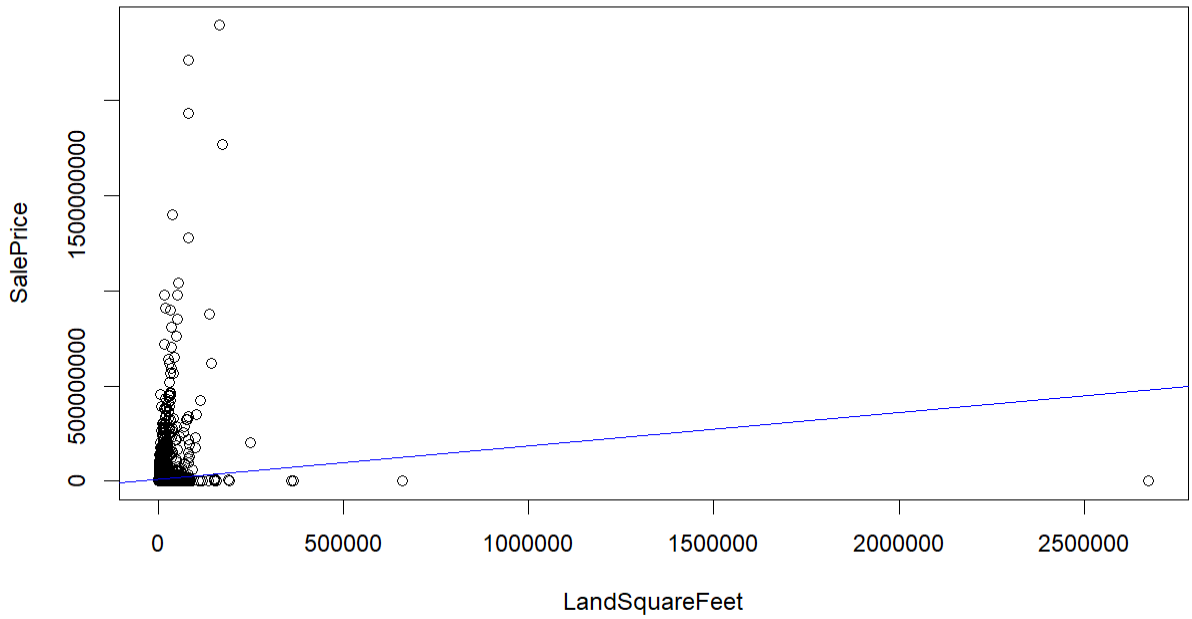
The regression plot for Sales Price vs. Gross Square Feet is:



The linear model is:

$$SalesPrice = 4837859.215 + 138.014 * GrossSquareFeet$$

The regression plot for Sales Price vs. Land Square Feet is:



The linear model is:

$$SalesPrice = 8109966.52 + 176.49 * LandSquareFeet$$

As we can see in the two plots above, the land square feet has more outliers than gross square feet. The concentration of points for land square feet is higher than for gross square feet, making the outliers a little more obvious. However, gross square feet has more distribution, so the points are further apart rather than centralized in one area. In terms of values, gross square feet was more proportional to the sales price than land square feet. This makes sense because gross square feet includes all floors in a building.

The linear model for Sales Price vs. Gross Square Feet and Land Square Feet is:
$SalesPrice = 6766244.443 + 228.564 * GrossSquareFeet + (-558.026) * LandSquareFeet$

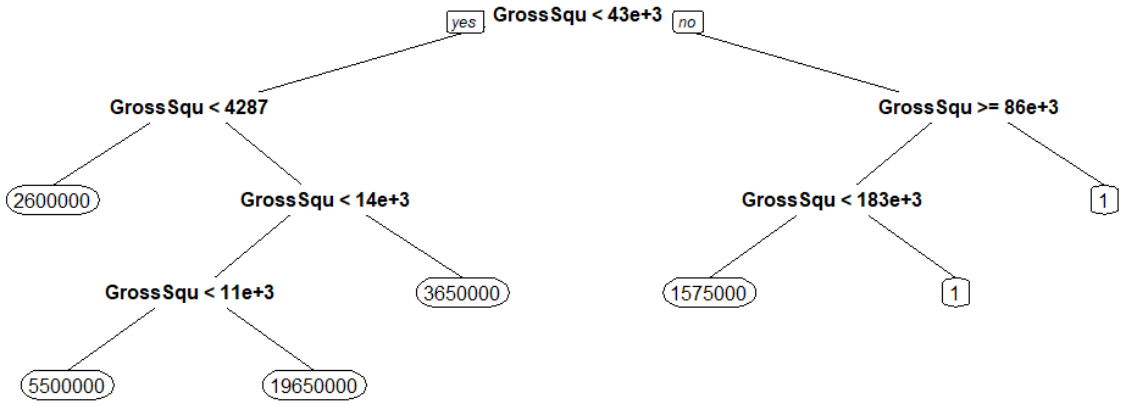
The summary for the above is:

```
Call:
lm(formula = SalePrice ~ GrossSquareFeet + LandSquareFeet, data = multi)

Residuals:
       Min         1Q     Median         3Q        Max
-557307014   -7305210   -4250967    2251352 2231379205

Coefficients:
                   Estimate  Std. Error t value            Pr(>|t|)
(Intercept)     6766244.443  477402.204   14.17 <0.0000000000000002 ***
GrossSquareFeet     228.564       4.786   47.76 <0.0000000000000002 ***
LandSquareFeet     -558.026      21.299  -26.20 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52190000 on 13048 degrees of freedom
Multiple R-squared:  0.1567,     Adjusted R-squared:  0.1566
F-statistic:  1212 on 2 and 13048 DF,  p-value: < 0.00000000000000022
```

From all three models, it is evident that gross square feet has a bigger impact on sales price than land square feet. As mentioned earlier, this is practical because gross square feet includes the area of all floors of a building rather than just the land. The models are all statistically significant because the p-value is less than 2.2e-16. In conclusion, the gross square feet affects the sales price of a residence more than land square feet.

d. Pick one or more models (these need not be restricted to the models you've learned so far [Decision Trees, KNN, K-Means, RandomForest…]) to explore the chosen data. Interpret the model fits and indicate significance. Describe any cleaning you had to do and why. Min. 5 sentences (2%)

The decision tree is as follows:



As visible in the image above, the decision tree starts at the top. If the gross square feet is less than 43,000, it goes to the left, otherwise it goes to the right. It continues on until it reaches a decision. For example, if the gross square feet of a specific property is 2,000 square feet, it is predicted to have a sales price of $2,600,000 in Manhattan. This seems reasonable solely because it is Manhattan and housing is expensive. The clean up that had to be done was omitting any NA values as well as 0s for any of the variables because it wouldn't make sense to have those.

2. For your chosen dataset:

a. Apply the model(s) to predict quantities of interest (that you choose). Describe (contingency table) or plot the predictions. Min. 2-3 sentences (4000-level 5%, 6000-level 3%)

I chose to use a multivariate regression model to analyze the effect that both gross square feet and year built have on the sales price, with sales price being the dependent variable and the other two as independent variables. I predicted that there will be a positive correlation between both year built and gross square feet. A building that was built more recently and with a larger gross area will increase the sales price. The multivariate regression model summary is below:

```
Call:
lm(formula = SalePrice ~ GrossSquareFeet + YearBuilt, data = d)

Residuals:
        Min          1Q      Median          3Q         Max
-1560223699    -9732520    -5671012    -1420495  2195358526

Coefficients:
                    Estimate   Std. Error t value             Pr(>|t|)
(Intercept)     135607380.50  32782487.10   4.137            0.0000356 ***
GrossSquareFeet       173.60         4.72  36.777 < 0.0000000000000002 ***
```

```
YearBuilt          -65361.95     16917.15  -3.864            0.000113 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65900000 on 8180 degrees of freedom
Multiple R-squared:  0.142,      Adjusted R-squared:  0.1417
F-statistic: 676.7 on 2 and 8180 DF,  p-value: < 0.00000000000000022
```

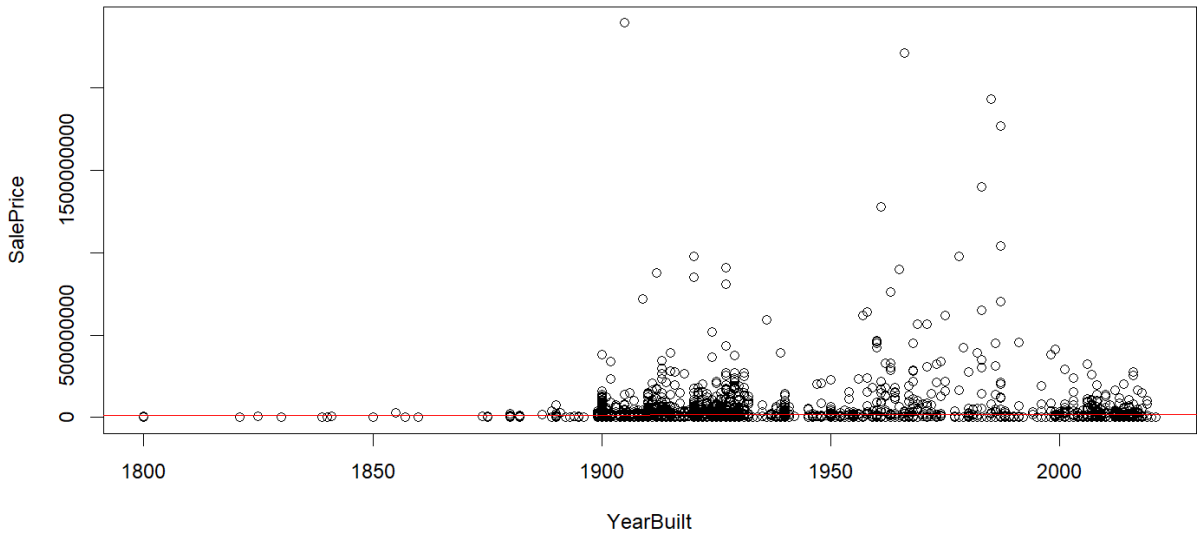
The equation is:

$$SalesPrice = 135607380.5008 + 173.5982 * GrossSquareFeet + (-65361.9487) * YearBuilt$$

b. Examine the fit(s). Perform a significance test that is suitable for the variables you are investigating and describe the results. Min. 2-3 sentences (4000-level 4%, 6000-level 3%)

As can be seen in the summary from part (a), the model has a p-value of less than 2.2*10^16, which shows that it is statistically significant. Unlike my prediction, the year built has a negative correlation with the sales price, meaning that an earlier year built would most likely have a higher price. This contradicts my prediction, as I thought that people would be more willing to buy a newer home that is more appealing. The sales price increases as the gross square feet increases, which was as expected.

If we isolate year built and sales price, the regression model is as follows:




```
Call:
lm(formula = SalePrice ~ YearBuilt, data = d)

Residuals:
```

```
       Min         1Q     Median         3Q        Max
 -16677256  -14435833  -11397055   -5523435 2382476108

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13386059   35115661  -0.381    0.703
YearBuilt       14914      18109   0.824    0.410

Residual standard error: 71140000 on 8181 degrees of freedom
Multiple R-squared:  8.291e-05,  Adjusted R-squared:  -3.932e-05
F-statistic: 0.6783 on 1 and 8181 DF,  p-value: 0.4102
```

The ab line is completely horizontal, meaning there is no significant relationship between year built and sales price. However, the p-value is 0.4102, showing that it is not very statistically significant. This is interesting because one would think that a newer property would lead to a higher sales price, but this is not the case. There is no direct correlation between year built and sales price, as we can see in the regression line.

c. Discuss any observations you had about the datasets/variables, other data in the dataset and/or your confidence in the result. Min 1-2 sentences (1%)

Overall, it is evident that the gross square feet has a positive correlation with sales price, so as the gross square feet increases, the sales price of the property also increases. The same goes for land square feet. I am pretty confident with this result because it is clear through the regression models that there is a direct positive correlation for both of these variables, with gross square feet having a slightly higher correlation. On the other hand, year built does not seem to have a direct negative or positive correlation if we look at the isolated model, however, the multivariate regression model suggests a slight negative correlation. Because of this, my confidence on this variable is low.