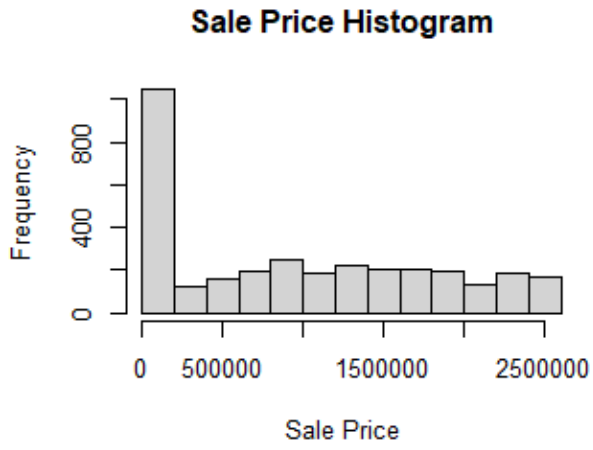
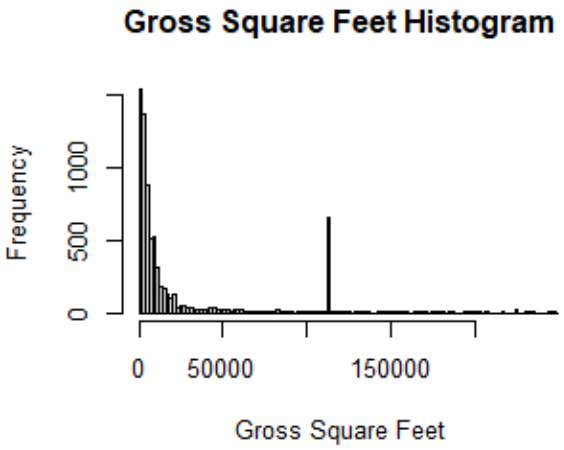
Tony Min
Data Analytics
November 3, 2023

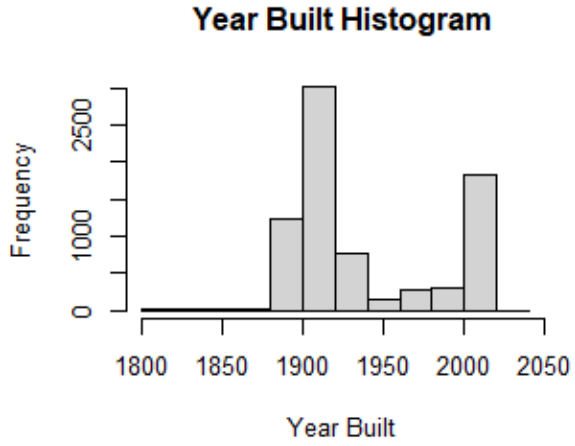
Assignment 4

1.

a. For the NYC Citywide Annualized Calendar Sales Update dataset I decided to pick Brooklyn as the borough of focus. I am deciding to look at what affects the sale price of. The variables that could have correlation to this from my own knowledge would be year built and the gross square feet. We first had to clean up some of the data within the dataset. For gross square feet I converted the value within that column into integers, for the year built column we got rid of null / missing values and got values that were greater than 0, this was also done for sale price. We can learn whether new properties sell for higher prices and or using gross square feet we could see if that affects price.







Looking at these histograms we see that the sale price of houses are usually below $500,000 while there are some more expensive properties that reach above $2,500,000 which could be an outlier, we will look into these data points in the next problem. Now on the square feet most properties are before 50,000 square feet. Finally for the year built we see that most properties were built after 1900. We show the summary statistic values in the table below.

	Minimum	1st Quarter	Median	Mean	3rd Quarter	Maximum
Sale Price	1.00	1.17e+06	3.6e+06	1.6e+07	9.2e+06	2.4e+09
Gross Square Feet	91	2592	6055	39658	20000	2400000
Year Built	1800	1910	1920	1941	1989	2021

b. Using the Cook's Distance we decided that the outliers would be 3 times greater than the mean, the outliers are shown below.

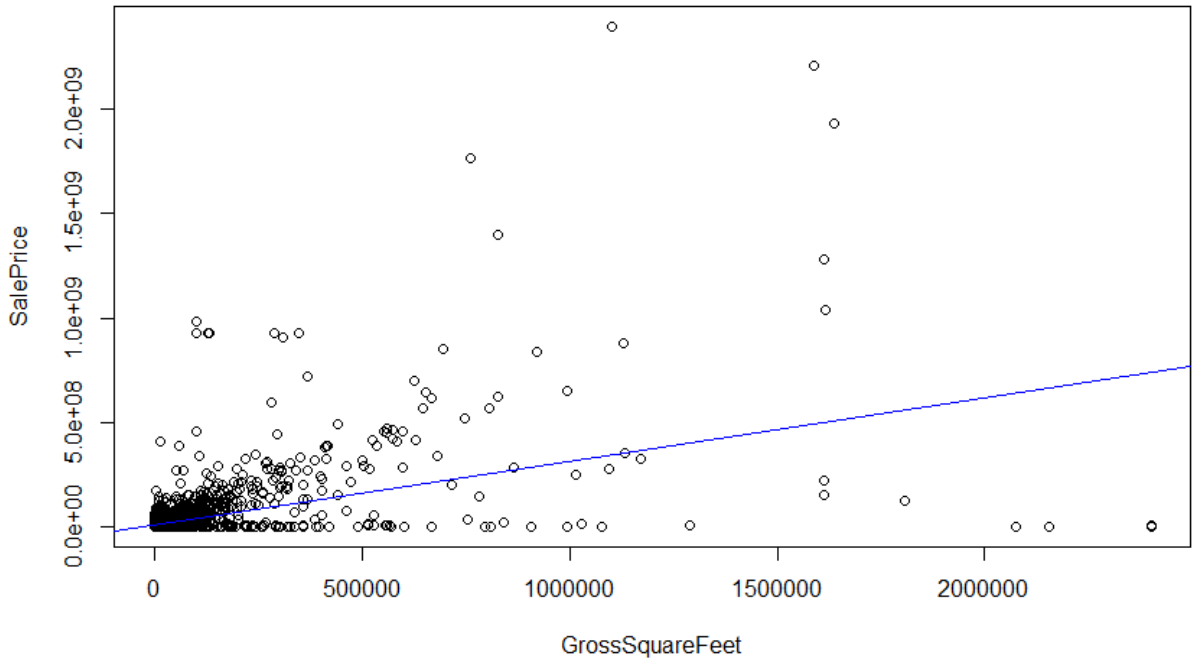
```
> nameofinfluential
 [1] "4"    "70"   "86"   "524"  "526"  "540"  "545"  "606"  "613"  "614"
[11] "964"  "969"  "1251" "1294" "1546" "1580" "1839" "1840" "1841" "1858"
[21] "1916" "2267" "2613" "2707" "2895" "3126" "3176" "3380" "3521" "3690"
[31] "3696" "3697" "3882" "4260" "4580" "4755" "4935" "5583" "5845" "5888"
[41] "6490" "6492" "6893" "6919" "6975" "7011" "7012" "7039" "7425" "7426"
[51] "7427" "7428" "7429"
```

There are a total of 53 outliers and this screenshot. The process of finding these outliers was to create a linear regression model with sales price as the dependent variable and gross square feet and year built as the independent variables. We had to create a new data frame that contained only these variables in order to make the model easily. After the model we use Cook's Distance function to calculate the Cook's Distance and as we said before we decide outliers would have a distance of 3 times greater than the mean.

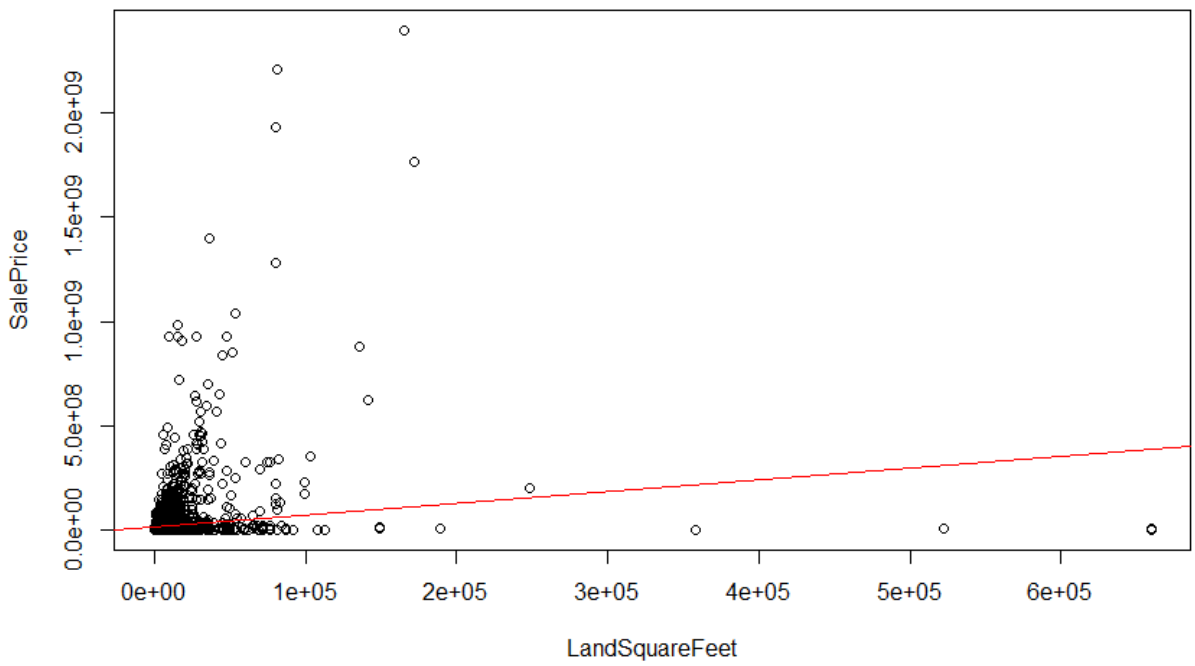
Now using the IQR to identify the outlier points we get 776 outliers, which is a lot more. The reasoning for this is how the IQR is calculated, we found the lower and upper bounds by subtracting 1.5 * the different between quartile 3 and quartile 1 from quartile 1, and then adding 1.5 * the difference from quartile 3 respectively. From question one we know that the majority of properties where prices from below $500,000 and we found that the first quarter was 1,170,000 the difference there is not 1.5 times the difference therefore the many of the properties were counted as outliers.

c. Conducting a Multivariate Regression we predicted Sales Price using Gross Square feet and or Land Square feet. We choose 3 different samples 1. Sales Price, Gross Square feet and Land Square feet 2. Sales Price and Gross Square

Feet 3. Sale Price and Land Square Feet. Looking at the first sample we used a linear model with Sales Price as the dependent variable and then Gross Square feet and Land Square feet as independent variables. We found 6336000 + 355.5 * GrossSquareFeet - 481.5 * LandSquareFeet was the linear model equation to find the sale price based on the independent variables. As we can see the Gross Square Feet has a positive effect in predicting Sale Price while Land Square feet has a negative effect. This model explains about 24.43% of the variability in Sale Price suggesting that other factors that are not included in the model may influence Sale Price. Furthermore, with a residual standard error of 65,590,000 the model may be sensitive to outliers.



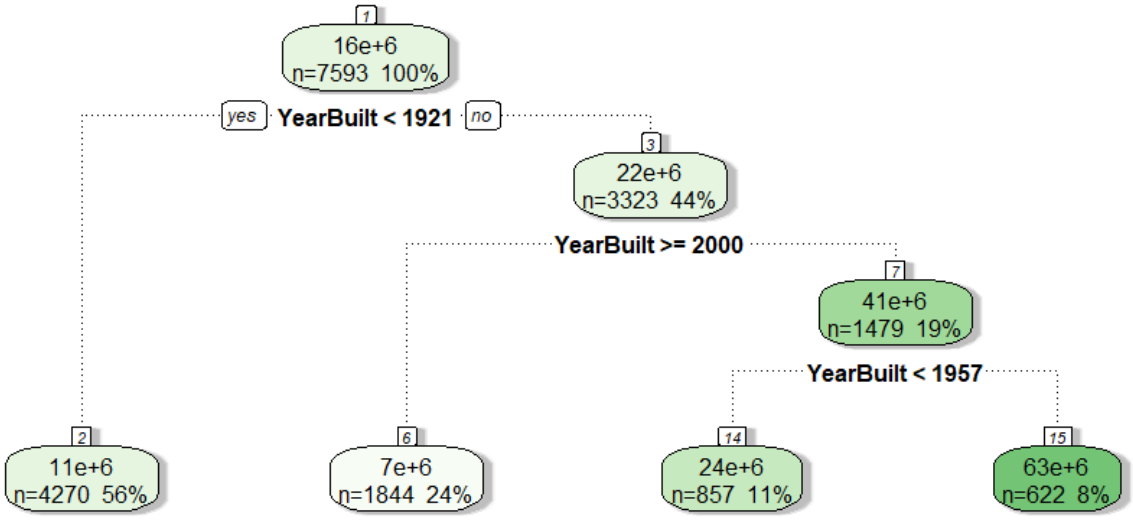
Above is the regression plot for SalePrice vs GrossSquareFeet with the linear model line. Now with the second model of Sales Price and Gross Square Feet we get a linear model equation of SalePrice = 3,512,00 + 306.9 * GrossSquareFeet. As we see again the GrossSquareFeet has a positive effect on SalePrice but comparing it to the other same we get a 23.07% of variability in SalePrice. This model again has a large residual standard error so this model may also be sensitive to potential outliers.



Above is the regression model for SalePrice vs. LandSquareFeet with the red line being the linear model equation of SalePrice = 10,110,000 + 565.2 * LandSquareFeet. This model predicts a very low amount of variability of SalePrice of only 2.68%

Comparing all three models we can see that the first model is the most accurate followed by the second and then the third. We see that the impact of Gross Square Feet is greater than the impact of Land Square Feet. This makes sense since gross square feet is all of the living space area instead of just the land that is on the property, meaning if there is more than one floor the the gross square feet would increase and be an accurate representation of the price of the property compared to the land square feet. Also for all models we had a p-value of 2.2e-16 meaning they were all statistically significant as well.

d. I have decided to create a decision tree for the Sale price based on the year the property was built, it is shown below.



Rattle 2023-Nov-02 20:49:59 Tony

As we see in this decision tree we can see how the decision tree is formed. From the top we see that if the property was built before 1921 with a sale price of 1.1000,000 on the second node down we see that any property built after 2000 with a sale price of 7,000,000. On the last layer we see that with a property built before 1957 the cost property cost is 24,000,000 while other properties that are built after 1957 that cost way more at a value of 63,000,000. I find it interesting how the older properties cost more than the newer properties, another factor to this could be how large these properties are since older properties may have been built larger while current properties are smaller. This could be due to the fact that in New York City its highly populated in a smaller area properties may want to take up less area in order to fit in another property, leading to more people being able to live in that area.

2.

a. I would choose a multivariate regression model to find how gross square feet and year built affects the sale price of these properties in Brooklyn. After creating this model we see there the equation comes out to be SalePrice = 263,300,000 + 317.2 * (GrossSquareFeet) - 134,100 * (YearBuilt). From this equation we see that GrossSquareFeet has a possible effect, increasing the SalePrice while the YearBuilt decreases the Sale Price. The prediction in conclusions shows that a property with more GrossSquareFeet would increase the price while a newer property leads to a decrease in price. This seemed expected since from the last question the decision tree showed how older properties had a high Sale Price.

b. In the summary statistics of the model we had a p-value of 2.2e-16 which indicates that this model was statistically significant due to a low p-value. From the results of this I found that the increase in GrossSquareFeet leads to an increase in SalePrice since usually the more living area a property has the greater the price. I know from our decision tree that a newer property leads to a lower price but to find out if that is actually true we run another regression model on the SalePrice as the dependent and YearBuilt as a independent and generate an equation of SalePrice = -62078264 + 40069 * (YearBuilt). With this model in the summary statistic the p-value was 0.04102 which is less than 0.05 therefore the model is also statistically significant therefore YearBuilt does have correlation in predicting the SalePrice.
c. From working through these models and observations we can come to a conclusion that out of the whole nyc Brooklyn dataset some of the variables that are statistically significant in modeling the SalePrice are GrossSquareFeet, LandSquareFeet and YearBuilt. For the first 2 variables we know there is a positive correlation while for YearBuilt there is a negative correlation. When the GrossSquareFeet and LandSquareFeet increase the SalePrice will increase while the opposite happens with YearBuilt. While it seemed to make sense that when you increase the square feet the price increased, it was surprising that year built would have a decrease in the price. The reason why this may be happening could be as we explained in the decision tree problem but to reiterate in the current years due to the high population in Brooklyn the properties are now being constructed in a way to hold the most people in an area instead of just building a big property. These smaller properties lead to a smaller gross square feet which as we know from the model leads to a decrease in sale price as well. I think from this analysis and from what we know about the current times it shows how newer properties could cost less than older properties.

3. In this study we created multiple linear regression models for SalePrice based on a combination of GrossSquareFeet, YearBuilt and LandSquareFeet. From these models we saw that all 3 of these variables were statistically significant in terms of predicting SalePrice. These findings aligned with how I thought these variables would affect the Sale Price due to the background knowledge I have in the housing prices and the living conditions in New York City. The model was suitable for seeing the relationship between variables and how they affect our dependent variable, SalePrice.
   However, it seemed that either the variables I chose or model weren't able to explain the variance in SalePrice, most of the model only explained up to 25% of it. This may be because there was another variable or a combination of variables which explain more of the variance or since the model used the normal dataset without removing the outliers. With a little more cleaning of the dataset, like removing outliers from the model we may be able to see a higher explained variance from the linear regression model. In

conclusion, using linear regression we found relations between GrossSquareFeet, LandSquareFeet, YearBuilt and how it affects SalePrice. I think this based exploration of the dataset worked for what we needed but without closely cleaning the dataset I would say we could not accurately predict the sale price of a property with our variables and model. To improve we should remove outliers and find more variables that could relate to Sale Price and rerun using a linear regression model or we could use a different model.