Property Assessment Values Analysis

Dany Hachem, Greg Cameron, Atlanta Liu

October 17, 2019

Contents

1 Introduction

The City of Calgary releases property assessment values on an annual basis. These assessments play an important role in determining the amount of property taxes that homeowners must pay to maintain city services. The city claims that it does its best to promote fairness and equity in its assessments by utilizing a sales comparison approach (Property Assessment 2019). This approach considers physical characteristics of the residential property, nearby community services, and surrounding property within the community. However, there is comparatively little mention of how external factors could influence the values of a property assessment. For this project, factors of interest that will be examined in relation to property assessment values include economic downturn (oil price), population growth, as well as crime rates.

Our report aims to further our understanding of how accurate these assessments are by looking at the extent to which each factor of interest can affect property assessments. For individuals who do not currently own any residential property, the insight gathered from our analysis should hopefully provide them with a better awareness of how important these factors are when determining whether one should purchase certain residential properties. In general, this analysis will provide a great starting point to discuss why the observed values of residential property assessments are seen to increase and decrease over the years.

The datasets for each factor will be obtained from Open Calgary. Using the data wrangling and visualization methods learned in class, we will be conducting exploratory data analysis of these factors on residential property assessments in Calgary to provide some insight into our guiding questions. All the datasets used are in comma separated value (csv) format and are up to date as of September 2019.

Topics to Investigate

The focus of this investigation will be to better understand how the property assessments are impacted by our two chosen factors: the economy and crime. When compiling the annual assessments, the city takes into account real estate data from the past 3 years and other factors including: the size of the property, recent renovations, the year it was built, and the neighbourhood. The goal is for the assessment to be as close as possible to the true market value of the property. This is challenging as the value of a property depends on a myriad of factors and not all of these can be built in to the assessments. While we hope to find these data useful in our investigations, we also realize that these are not true market values. To address how the economy and crime have impacted property assessments, we will compare subsets of the population to answer questions about the changes in property values. For example, were the assessed values of the large downtown office buildings more directly impacted by the decline in oil prices than suburban homes? Are the assessed values in inner city neighbourhoods more influenced by crime rate than outer suburbs? Does crime occur more in communities with lower assessment values? Is there a correlation between the change of crime rates and change of assessment values over the years?

2 Data Sources

Most of the datasets used for this project originated from the Open Calgary website (data.calgary.ca). The main dataset used was the property assessment dataset. It consists of over 6 million rows of csv formatted data containing property assessments for all properties in the City of Calgary for the last 15 years (2005 to 2019). We merged this with several other datasets, including the crime and disorder statistics and population datasets. We also obtained oil price information from investing.com. The data were cleaned and wrangled to remove any extraneous information as well removing records with partial data. Median property assessment values were computed per community per year, which form the basis for all of our statistical work for this project.

3 Bootstrap Mean μ and Standard Deviation σ Analysis

Before diving into statistical analysis, it is always best to look at how are out data is distributed. To that extent, we plotted the data for Assessment Values, Crime Counts, as well as Population for the years 2012 and 2017. Given that we have around 200 data points, we would expect the Central Limit Theorem to kick in which states that the sampling distribution will be well modelled by a Normal distribution.

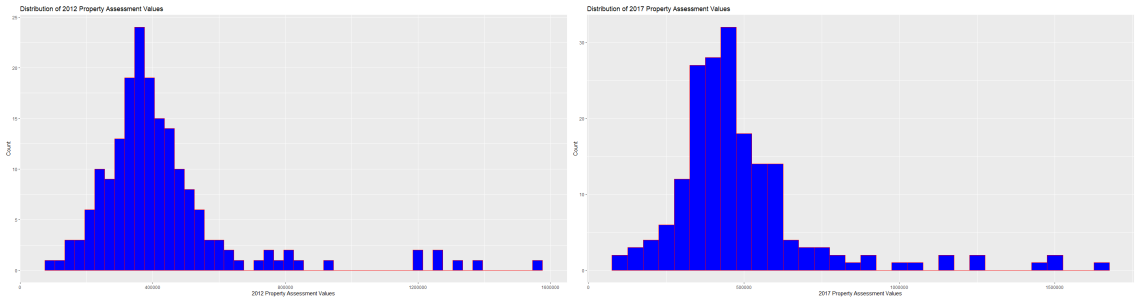


Figure 1: Distribution of Property Assessment Values in 2012 and 2017

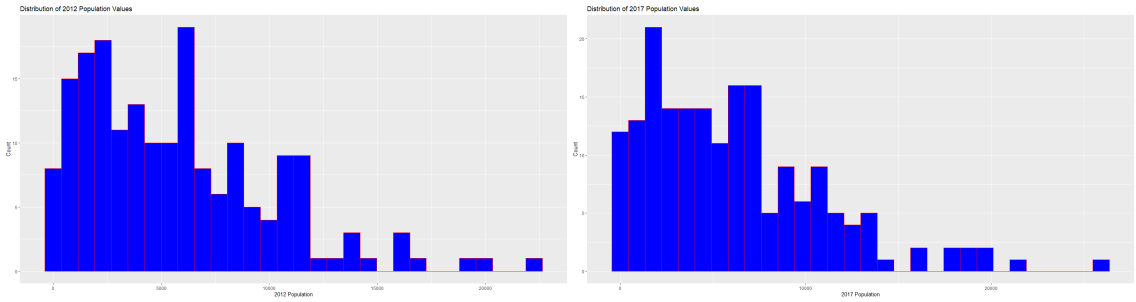


Figure 2: Distribution of Population Values in 2012 and 2017

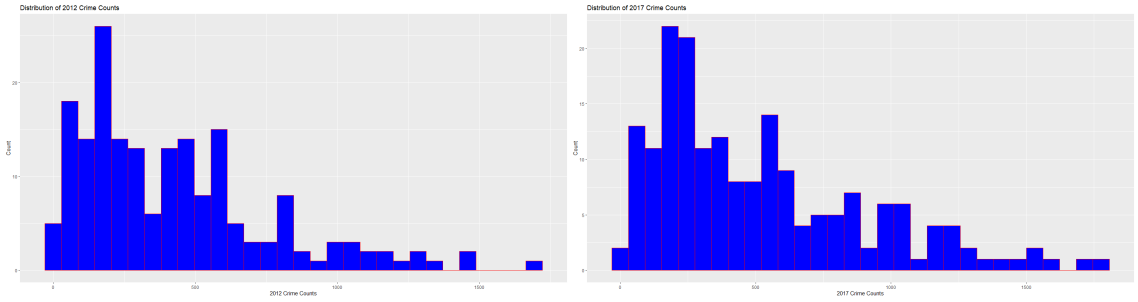
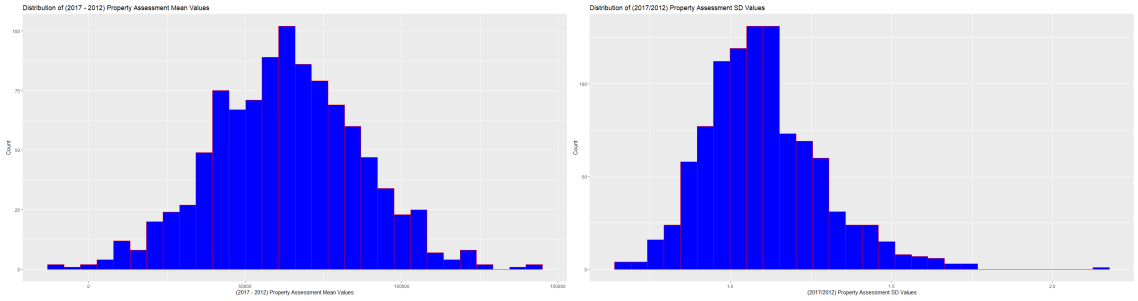


Figure 3: Distribution of Crime Counts in 2012 and 2017

As can be seen clearly, only Property Assessment Values follow a Normal distribution, but even those values aren't perfectly normal. This justifies our use of the bootstrap method to estimate mean difference intervals, standard deviation ratio intervals, and confidence intervals. Confidence intervals were computed in R using either the "qdata" command or the "quantile" command.



(a) $\mu_{2017} - \mu_{2012}$ (b) $\frac{\sigma_{2017}}{\sigma_{2012}}$

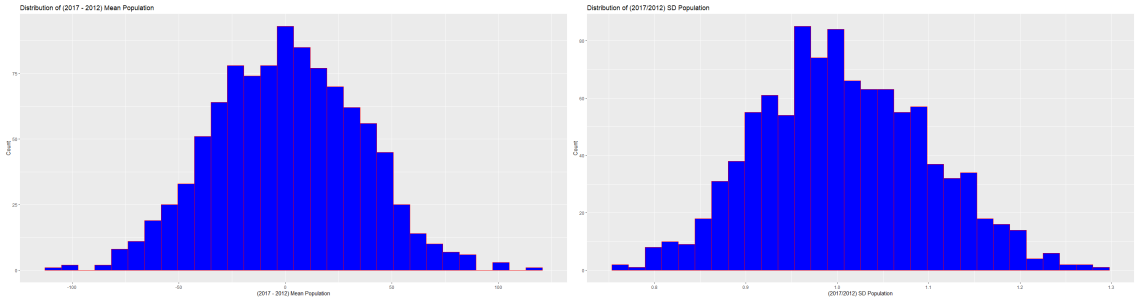
Figure 4: Assessment Values Bootstrap statistics

$$\text{A 95\% confidence interval: } 17644.09 \leq \mu_{2017} - \mu_{2012} \leq 107853.89$$

This C.I. suggests that the change in Property Assessment Values from 2012 to 2017 is somewhere between those two values, with 95% confidence. Since this interval is positive and does not capture zero, we can conclude that, statistically, the average Property Assessment increased by at least 17,644CAD and at most by 107,854CAD.

$$\text{A 95\% confidence interval: } 0.7969 \leq \frac{\sigma_{2017}}{\sigma_{2012}} \leq 1.5181$$

This C.I. suggests that, statistically, the standard deviation of Property Assessment is the same in 2017 as it was in 2012. This is because the ratio interval captures one (1).



(a) $\mu_{2017} - \mu_{2012}$ (b) $\frac{\sigma_{2017}}{\sigma_{2012}}$

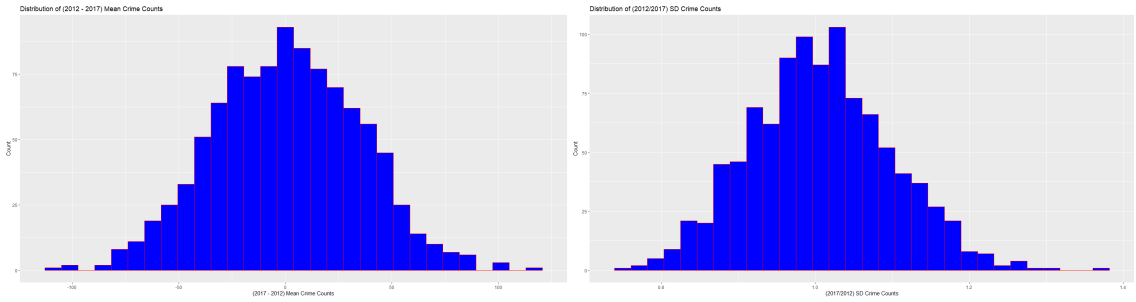
Figure 5: Population Bootstrap statistics

A 95% confidence interval: -64.72595 $\leq \mu_{2017} - \mu_{2012} \leq$ 66.58527

This C.I. suggests that the change in average Population from 2012 to 2017 is somewhere between those two values, with 95% confidence. Since the change interval captures zero (0), statistically speaking, we can say that the average population did not change from 2012 to 2017.

A 95% confidence interval: 0.8317 $\leq \frac{\sigma_{2017}}{\sigma_{2012}} \leq$ 1.1948

This C.I. suggests that, statistically, the standard deviation of Population is the same in 2017 as it was in 2012. This is because the ratio interval captures one (1).



(a) $\mu_{2017} - \mu_{2012}$ (b) $\frac{\sigma_{2017}}{\sigma_{2012}}$

Figure 6: Crime Count Bootstrap statistics

A 95% confidence interval: -64.7259 $\leq \mu_{2017} - \mu_{2012} \leq$ 66.5852

This C.I. suggests that the change in Crime Counts from 2012 to 2017 is somewhere between those two values, with 95% confidence. Since the change interval captures zero (0), we can statistically say that the crime counts did not change from 2012 to 2017.

A 95% confidence interval: $0.8329 \leq \frac{\sigma_{2017}}{\sigma_{2012}} \leq 1.1887$

Same as the previous two intervals, this C.I. suggests that, statistically, the standard deviation of Crime Counts is the same in 2017 as it was in 2012. This is because the ratio interval captures one (1).

To conclude this part of the analysis, we can say that based on the computed bootstrap statistics, the average Assessment Value increased from 2012 to 2017, while maintaining the same standard deviation. In terms of Population, we conclude that there is either no change, or a very slight increase from 2012 to 2017. And finally for Crime Counts, we can conclude that there is either a very slight increase in crimes per year from 2012 to 2019, or none at all. Both of these factors also maintained a constant spread from their respective average values from 2012 to 2019.

4 Bivariate Analysis of Oil Price and Property Assessment Values

4.1 Downtown Office Towers vs WTI Oil Price

We are investigating the relationship between the WTI oil price and two types of Calgary property assessments - the median assessment value of the large downtown office buildings and the median assessment value of all residential properties in Calgary. First, can it be explained with a linear relationship? Second, it appears that there is a delayed response to the prices for both types of assessments. Does the r^2 value improve when we delay the oil price by a number of years. What number provides the highest r^2 value? With the optimal shift in years, can the data be explained with a linear relationship?

The figure below shows three trends plotted for the years 2005 to 2019. The black line is the West Texas Intermediate (WTI) Oil Price in USD. The red line is the median assessed value of all office towers in Calgary's downtown core with a value of over $100 million (CAD) as of 2014. The blue line shows the median property values of all residential properties in Calgary. These data were cleaned and wrangled from the Property Assessment data from the Open Calgary website as part of our DATA 601 project.

We will first focus on the relationship between the Oil Price and the median assessed value of the office towers. The hypothesis for this study is:

H_0 - The relationship can not be explained by a linear function ($B = 0$)
H_A - The relationship can be explained by a linear function ($B \neq 0$)

The first step is to make a data frame in r of the Oil price and the median assessed value of the office towers and plot the results. The figure below shows a scatter plot of the data along with the line of best fit.

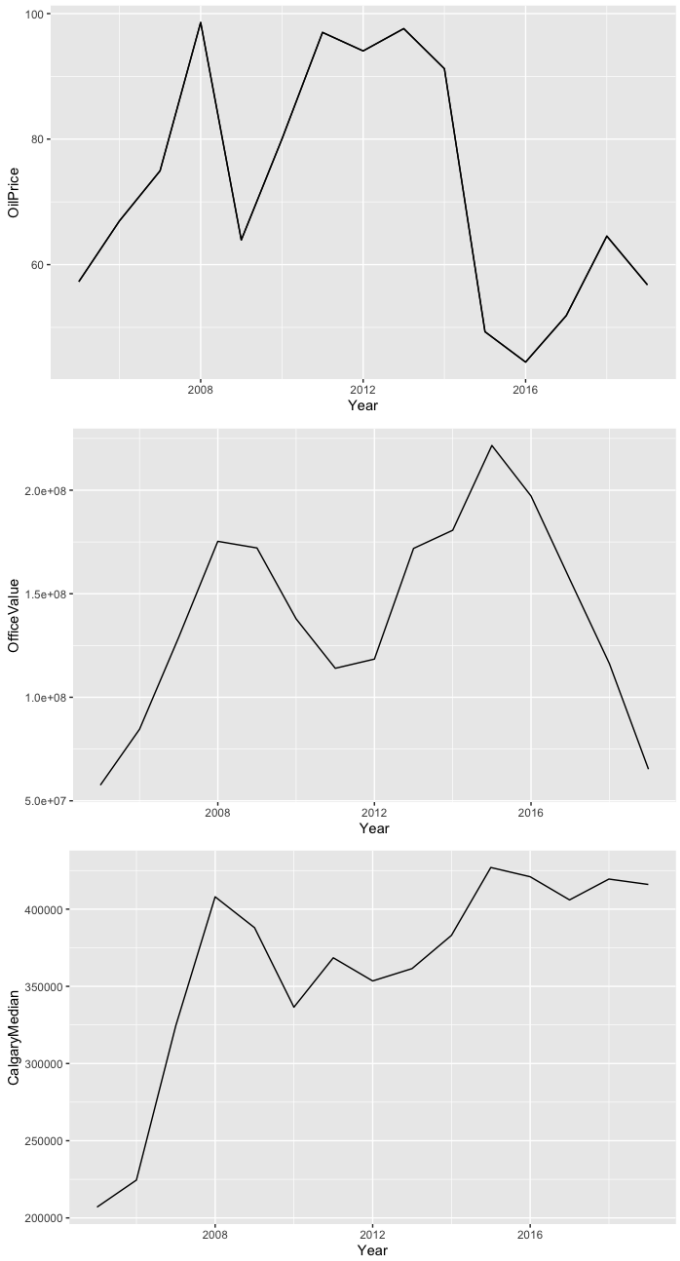


Figure 7: Oil Price, Median Office Tower, and Median Residential Assessment value in Calgary from 2005 to 2019.

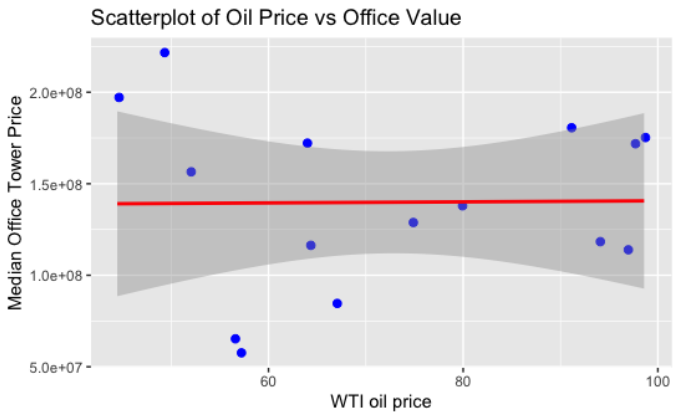


Figure 8: Scatter plot of WTI oil price vs median assessed value of Calgary downtown office towers.

Looking at the data, it does not appear that there is a linear relationship between the oil price and the median assessed values of the downtown office towers. To confirm this, we first compute the linear model. The results from the lm function in r give:

$$Y_{OfficeAssessment} = 137786901 + (28870)X_{OilPrice}$$

We then computed the p-value using the aov function, obtaining a value of 0.967. With this high p-value, we reject the null hypothesis. The assessment of the office towers can not be expressed as a linear function of Oil price. Computing the correlation using the cor function, we get a correlation value of 0.0115419, giving us an r^2 value of 0.0001332155. These values are not surprising given the data we observed on the scatter plot.

While the results above were expected given the data, we do think that the oil price should correlate with the office tower assessments as the tenants' income relies directly on the price they get for their commodities. However, there would most likely be a time delay to the correlation for a few different reasons. First, the assessment values are computed based on the year before they are released. Second, the market forces of lower corporate revenues would take some time to impact the property values. To evaluate this, we modified our data by delaying the oil price in yearly increments for 4 years. For example, when we delay the oil price by one year, we are now comparing the assessment value of 2015 with the oil price of 2014. The figures below show the equivalent scatter plots to figure 2 by delaying the oil price by 1, 2, 3, and 4 years.

We computed r^2 values all of the shifted datasets to estimate the amount of delay in years that led to the best correlation between the oil price and office tower assessment. The figure below shows the r^2 for varying amounts of shift.

From the figure, we see that the highest value for r^2 occurs when the oil price is delayed by 2 years. The linear model was recomputed for this dataset, giving:

$$Y_{OfficeAssessment} = 22525549 + (1653469)X_{OilPrice}$$

To re-test the hypothesis for these shifted data, we compute the p-value and get a value p=0.00136.

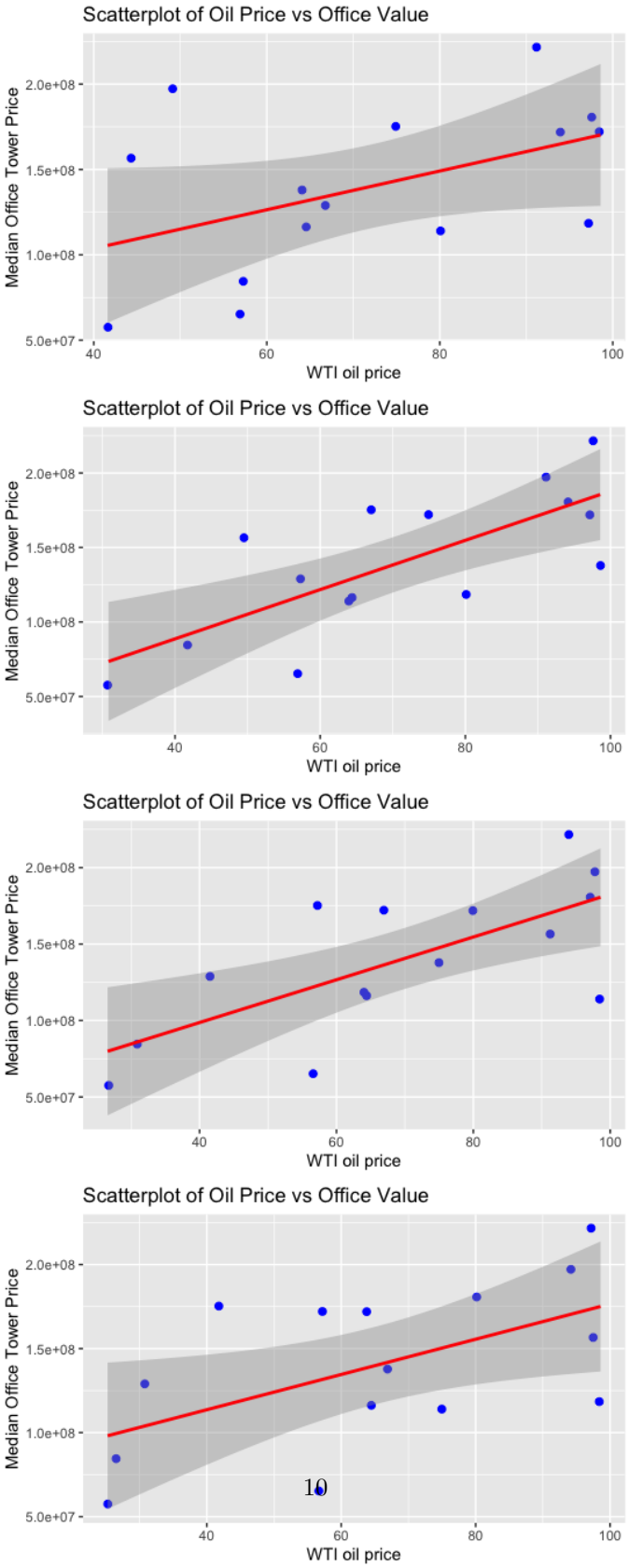


Figure 9: Scatter plots of WTI oil price vs median assessed value of Calgary downtown office towers. The oil price was shifted by (from top to bottom) 1, 2, 3, 4, years

Figure 10: plot of r^2 values for varying delay of oil prices

For this model, we reject the null hypothesis. The property assessments of the downtown office towers can be linearly modeled by the oil price when incorporating a 2 year time shift.

For this model to be valid, it must satisfy two conditions:

1. The y-variable, or commonly know as the response variable, is Normally distributed with a mean μ and standard deviation of σ (normality of the residuals).

2. For each distinct value of the x-variable (the predictor variable), the y-variable has the same standard deviation σ (homoscedasticity).

We test for the first condition by generating a normal probability plot of the residuals in R. We test for the second condition by running the plot of fits to residuals. The plots of both of these tests are shown below. The results show that the model satisfies both of the conditions.

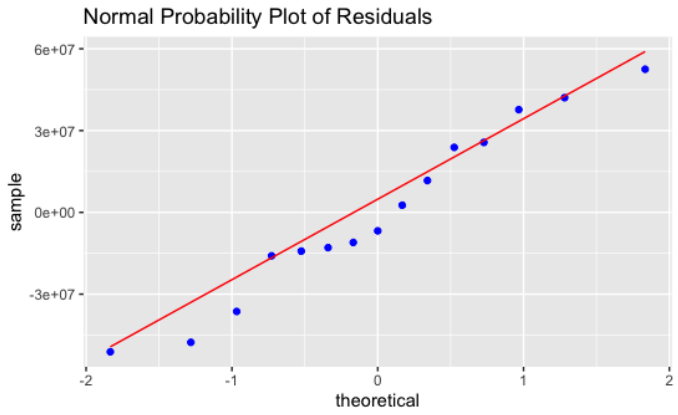


Figure 11: normal probability plot of the residuals

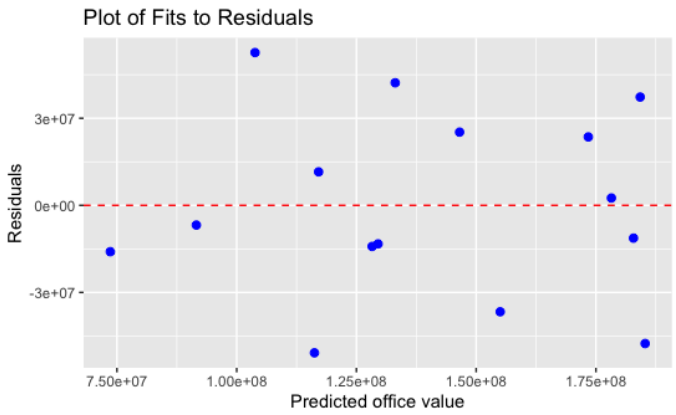


Figure 12: Plot of fits to the residuals

4.2 Residential vs WTI Oil Price

From the previous example we have observed that Calgary's downtown office towers can be modeled as a linear function of WTI oil prices. Is the same true for the Calgary's residential properties? Does the same time-delay of 2 years provide the strongest correlation?

The hypothesis is the same as for the previous example:

H_0 - The relationship can not be explained by a linear function ($B = 0$)
H_A - The relationship can be explained by a linear function ($B \neq 0$)

Figure 2 shows a scatter plot of the data along with the line of best fit.

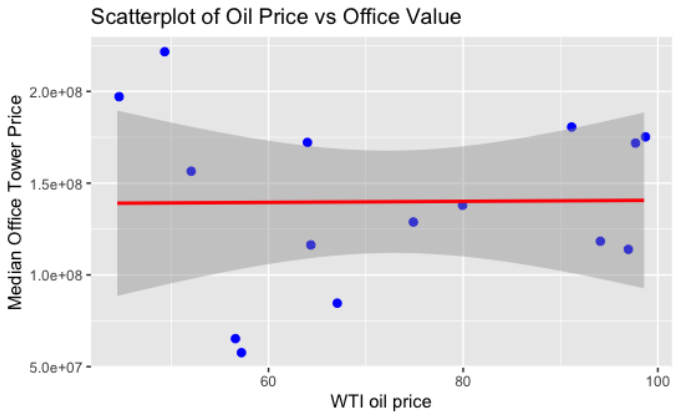


Figure 13: Scatter plot of WTI oil price vs median assessed value of Calgary residential properties.

Looking at the data, it does not appear that there is a linear relationship between the oil price and the median assessed values of the residential properties. To confirm this, we first com-

pute the linear model. The results from the lm function in r give:

$$Y_{ResidentialAssessment} = 379843.1 + (-232.6)X_{OilPrice}$$

We then computed the p-value using the aov function, obtaining a value of 0.967. With this high p-value, we reject the null hypothesis. The assessment of the residential properties can not be expressed as a linear function of oil price. Computing the correlation using the cor function, we get a correlation value of 0.0115419, giving us an r^2 value of 0.0001332155. These values are not surprising given the data we observed on the scatter plot.

While the results above were expected given the data, we would like to test if, as with the downtown office towers, there is a delayed relationship between the Calgary residential assessment values and the oil price. We follow the same method as above, delaying the oil price by 1, 2, 3, and 4 years, creating scatter plots and computing r^2 values.

For the residential property assessments, the highest value for r^2 occurs when the oil price is delayed by 2 years. The linear model was recomputed for this dataset, giving:

$$Y_{ResidentialAssessment} = 223143 + (2013)X_{OilPrice}$$

To re-test the hypothesis for these shifted data, we compute the p-value and get a value p=0.00136. For this model, we reject the null hypothesis. The property assessments of Calgary residential properties can be linearly modeled by the oil price when incorporating a 2 year time shift.

We also test this model for the conditions of normality of the residuals and homoscedasticity by generating the same plots as for the last example. The plots of both of these tests are shown below. The results show that the model satisfies both of the conditions.

The results of these two tests show that the property assessment values in Calgary can be linearly modeled using the price of oil. There is a delayed response to oil price changes due to market inefficiencies. Also, it shows that there is less delay between the price of oil and the downtown office towers assessments than between the price of oil and the residential assessments. We think these results make sense as it is the corporations in downtown that will feel the effects of lower commodity prices more quickly than the average residential property owner.

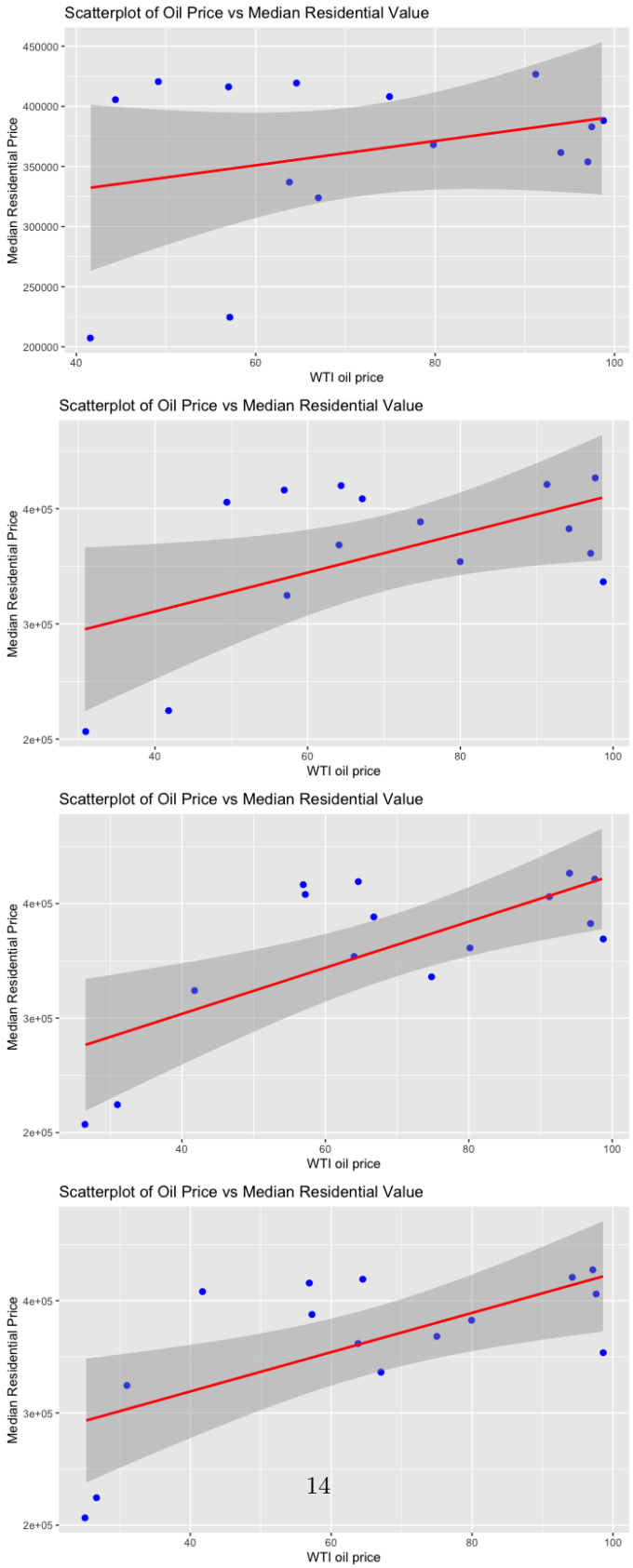


Figure 14: Scatter plots of WTI oil price vs median assessed value of Calgary residential properties. The oil price was shifted by (from top to bottom) 1, 2, 3, 4, years

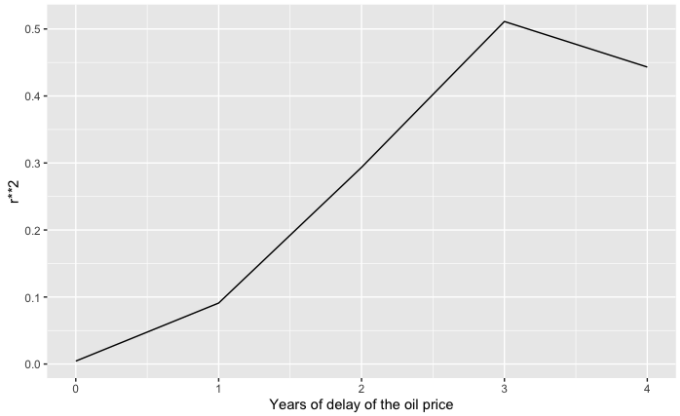


Figure 15: plot of r^2 values for varying delay of oil prices

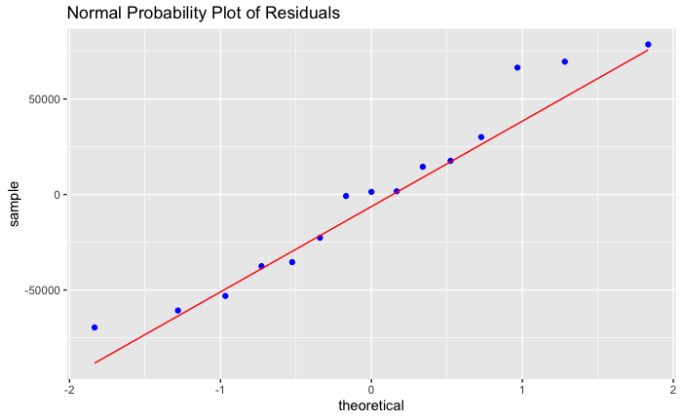


Figure 16: normal probability plot of the residuals

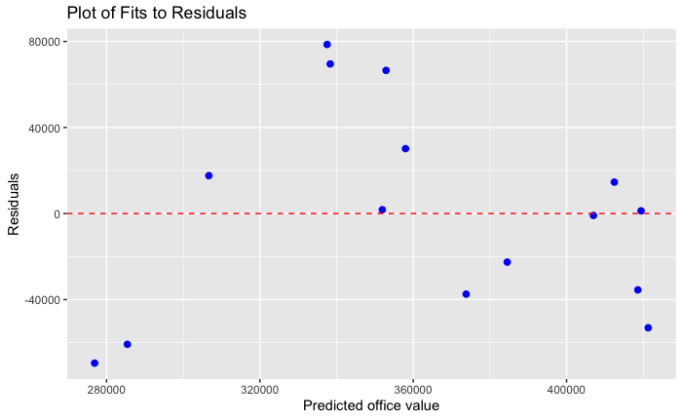


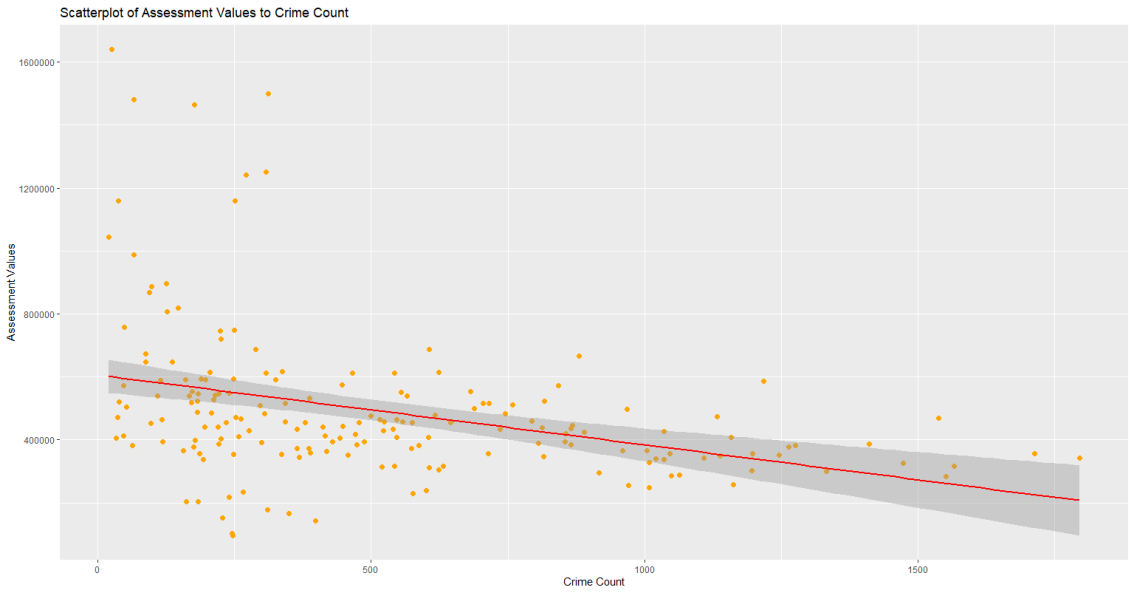
Figure 17: Plot of fits to the residuals

5 Bivariate Linear Modeling

5.1 Correlating Assessment Values with Crime for 2017

After cleaning and wrangling the datasets in Python, a csv file that contains values for Property Assessments, Population and Crime count was exported and used in R. Some communities were filtered out however, these being 'Beltline', 'Forest Lawn', 'Downtown Commerical Core', and 'Sage Hill'. The reason behind this is that 'Beltline', 'Forest Lawn' and 'Downtown Commerical Core' were extreme outliers in terms of crime counts, and 'Sage Hill' was an extreme outlier in terms of Assessment Values.

Figure 18: Assessment Values vs Crime Linear Model



First, a correlation factor was computed and turned out to have a value of r = *-0.3611*. As can be seen for the figure above, there seems to be an inverse relationship between assessment values and crime counts, where an increase in crime counts leads to decreased assessment values.

Based on our linear model, the relationship between Assessment Values and Crime Count can be modelled by:

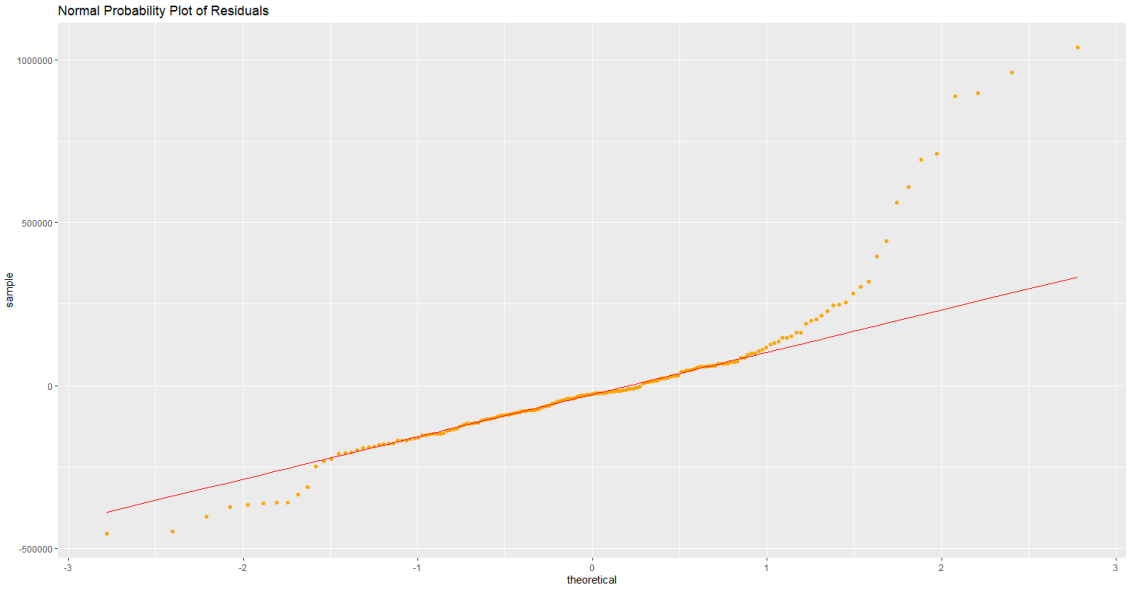
$$Y_{Assess} = 606286.7 + (-221.9)X_{CrimeCount}$$

Testing Conditions

Normality of Residuals

For the linear model to be valid, the Assessment Values must be normally distributed. To test this condition, Quantile-Quantile plots were used.

Figure 19: Quantile-Quantile Plot.

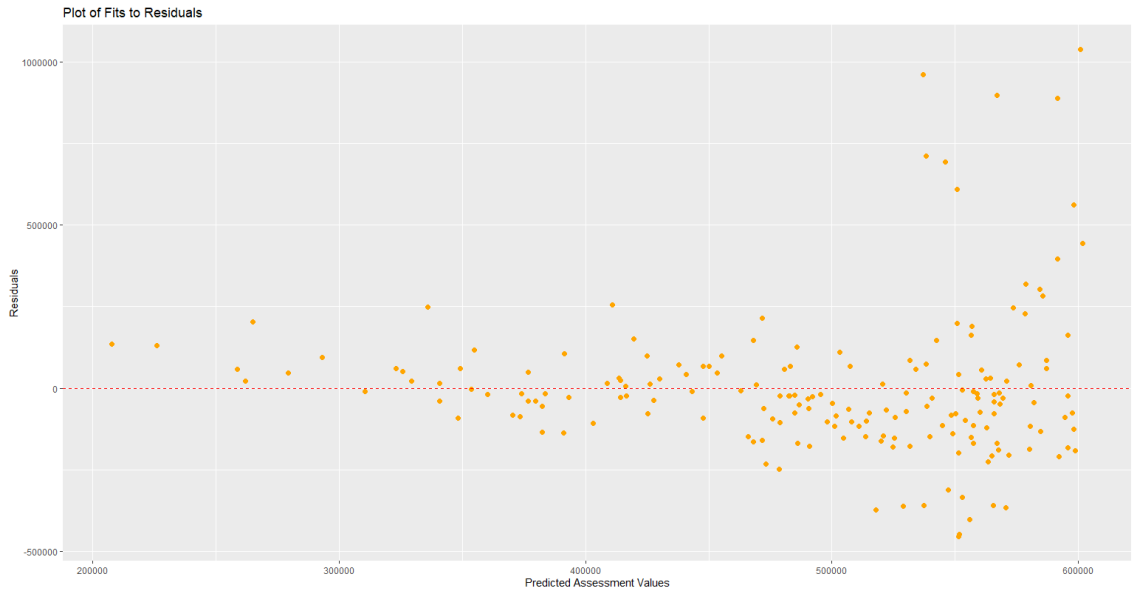


Although most of the data does lie on the straight line, it is apparent that the tail ends of the distribution are not adequately normal. However this deviation is not deemed too drastic given the proportion of data points that do conform to normality.

Homoscedasticity

The other condition that must hold is the stability of the variance of Assessment Values, otherwise known as the homogeneity of variance. To check for this, we plot the fitted values and residuals on a scatter plot and check the spread.

Figure 20: Plot of Fits to Residual Plot.



Disregarding a few outliers on the top right, it can be said that the condition of homoscedasticity is satisfied seeing how the residuals are scattered around zero and bounded equally above and below.

Following that, the coefficient of determination was computed in R, and turned out to have a value of r^2=*0.13*. This value gives the percentage of the variation of the Assessment Values that can be accounted for by their linear dependency on Crime Counts.

Statistical Significance of the Linear Model

To test for the overall linear appropriateness of the model, the following hypothesis were proposed:

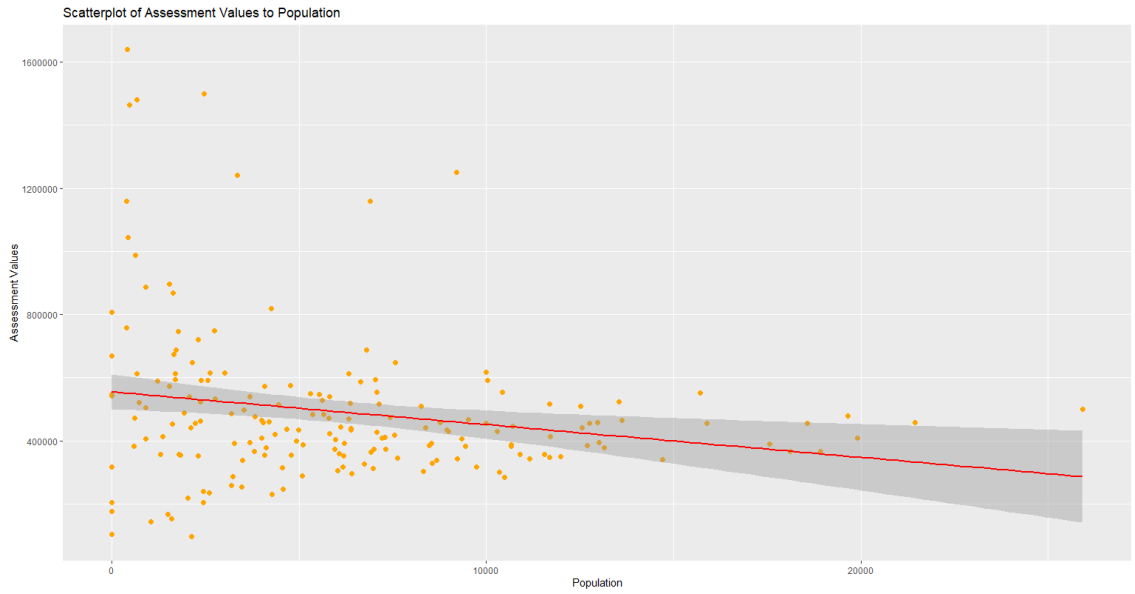
H_0: Assessment Values can not be expressed as a linear function of Crime Counts (B = 0)
H_A: Assessment Values can be expressed as a linear function of Crime Counts (B $\neq$ 0)

Using R, an F-test was conducted to determine the value of the F-statistic and the p-value. It turned out $F_{1,182}$ = *27.45* and *p-value* = *0.00000044*. This is a statistically significant result that allows us to reject the Null Hypothesis based on the fact that the *p-value* < *0.05*.

Based on that, we can say that the Assessment values can indeed be expressed as a linear function of Crime Counts.

5.2 Correlating Assessment Values with Population for 2017

Figure 21: Assessment Values vs Population Linear Model



First, a correlation factor was computed and turned out to have a value of r = *-0.2*. As can be seen for the figure above, there seems to be an inverse relationship between assessment values and population, where an increase in population leads to decreased assessment values.

Based on our linear model, the relationship between Assessment Values and Population can be modelled by:

$$Y_{Assess} = 554954.59 + (-10.37)X_{Population}$$

Testing Conditions

Normality of Residuals

For the linear model to be valid, the Assessment Values must be normally distributed. To test this condition, Quantile-Quantile plots were used.

Although most of the data does lie on the straight line, it is apparent that the tail ends of the distribution are not adequately normal. It is noticeably more deviating from normal compared to the crime counts, however this deviation is not deemed too drastic given the proportion of data points that do conform to normality.

Homoscedasticity

The other condition that must hold is the stability of the variance of Assessment Values, otherwise known as the homogeneity of variance. To check for this, we plot the fitted values and

residuals on a scatter plot and check the spread. Disregarding a few outliers on the top right, it can be said that the condition of homoscedasticity is satisfied seeing how the residuals are scattered around zero and bounded equally above and below.

Figure 22: Quantile-Quantile Plot.

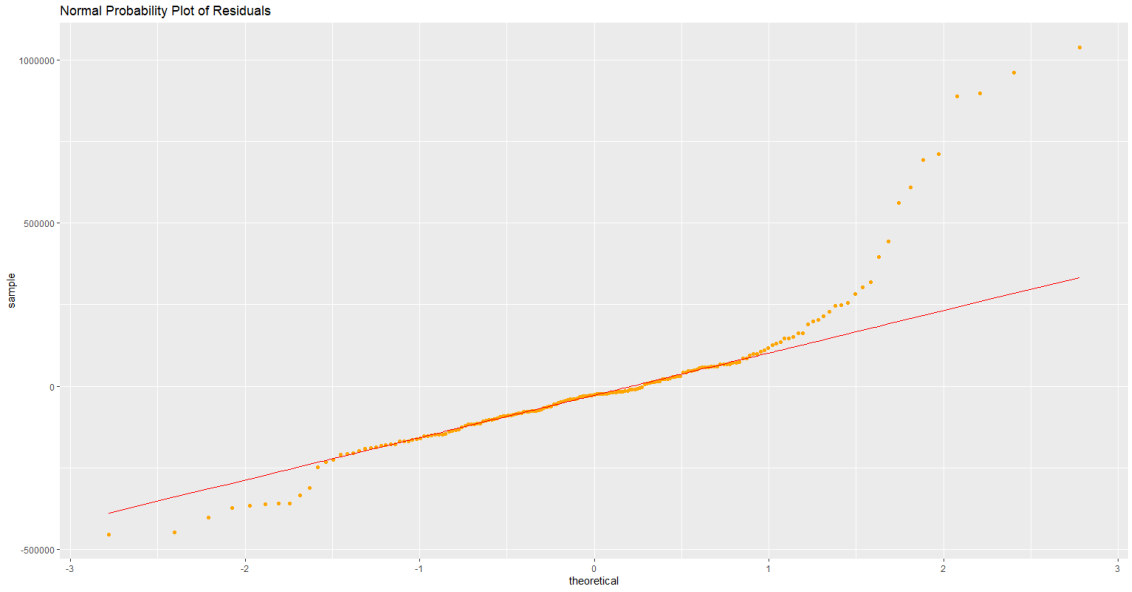
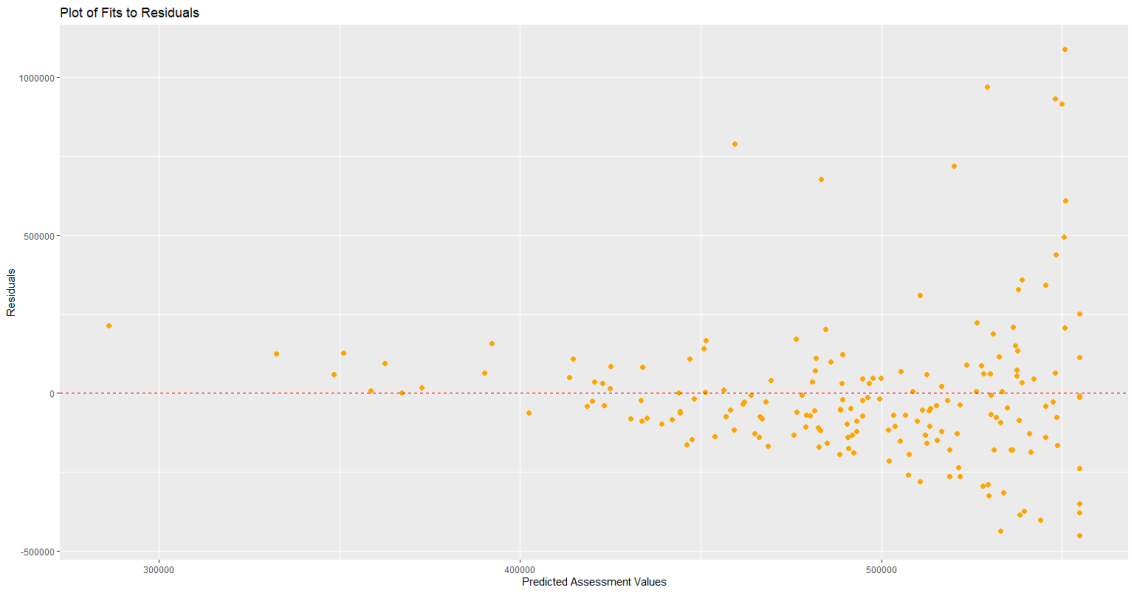


Figure 23: Plot of Fits to Residual Plot.



Following that, the coefficient of determination was computed in R, and turned out to have a value of r^2=*0.043*. This value gives the percentage of the variation of the Assessment Values that can be accounted for by their linear dependency on Population.

Statistical Significance of the Linear Model

To test for the overall linear appropriateness of the model, the following hypothesis were proposed:

H_0: Assessment Values can not be expressed as a linear function of Population (B = 0)
H_A: Assessment Values can be expressed as a linear function of Population (B $\neq$ 0)

Using R, an F-test was conducted to determine the value of the F-statistic and the p-value. It turned out $F_{1,182}$ = *8.254* and *p-value = 0.00455*. This is a statistically significant result that allows us to reject the Null Hypothesis based on the fact that the *p-value* $<$ *0.05*.

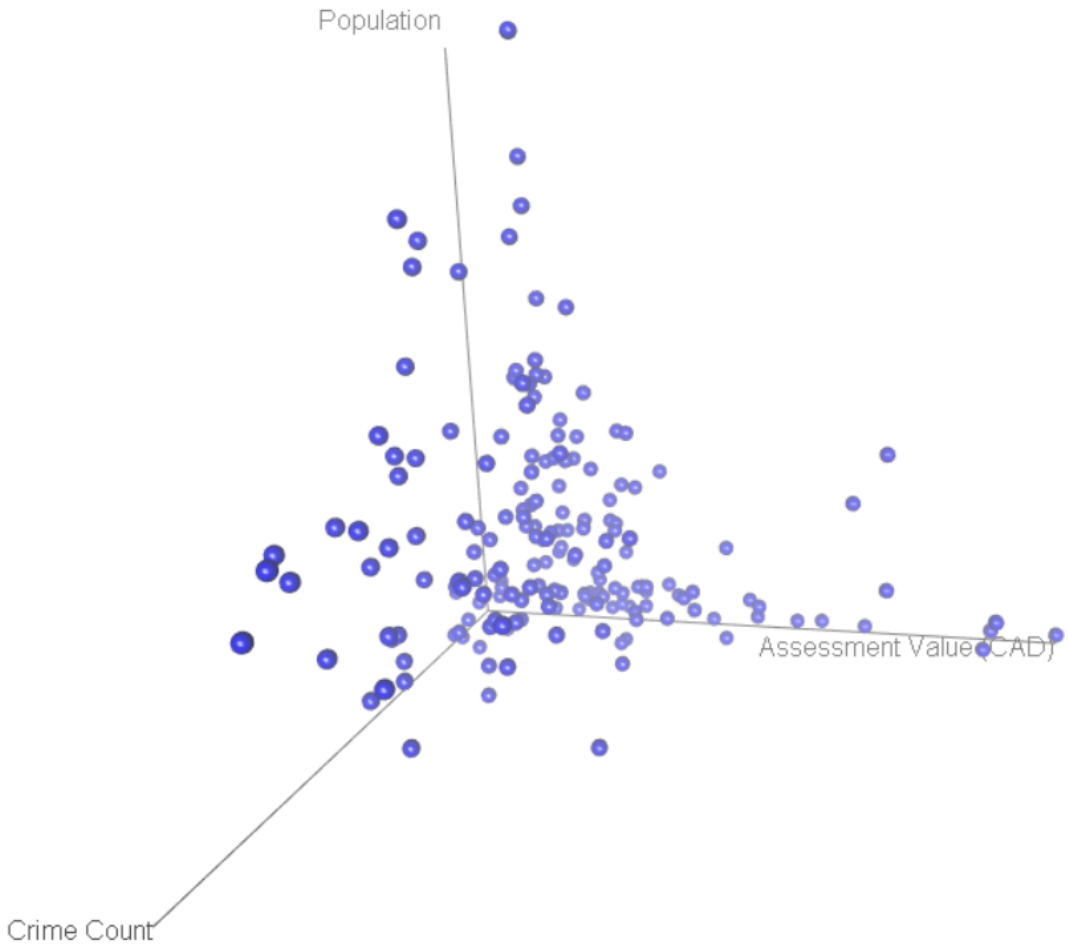
Based on that, we can say that the Assessment values can indeed be expressed as a linear function of Population.

5.3 Correlating Assessment Values with both Crime and Population for 2017

For this part of the analysis, we tried to see if Assessment Values can be modelled as a linear function of both Population and Crime Counts at the same time in the form of

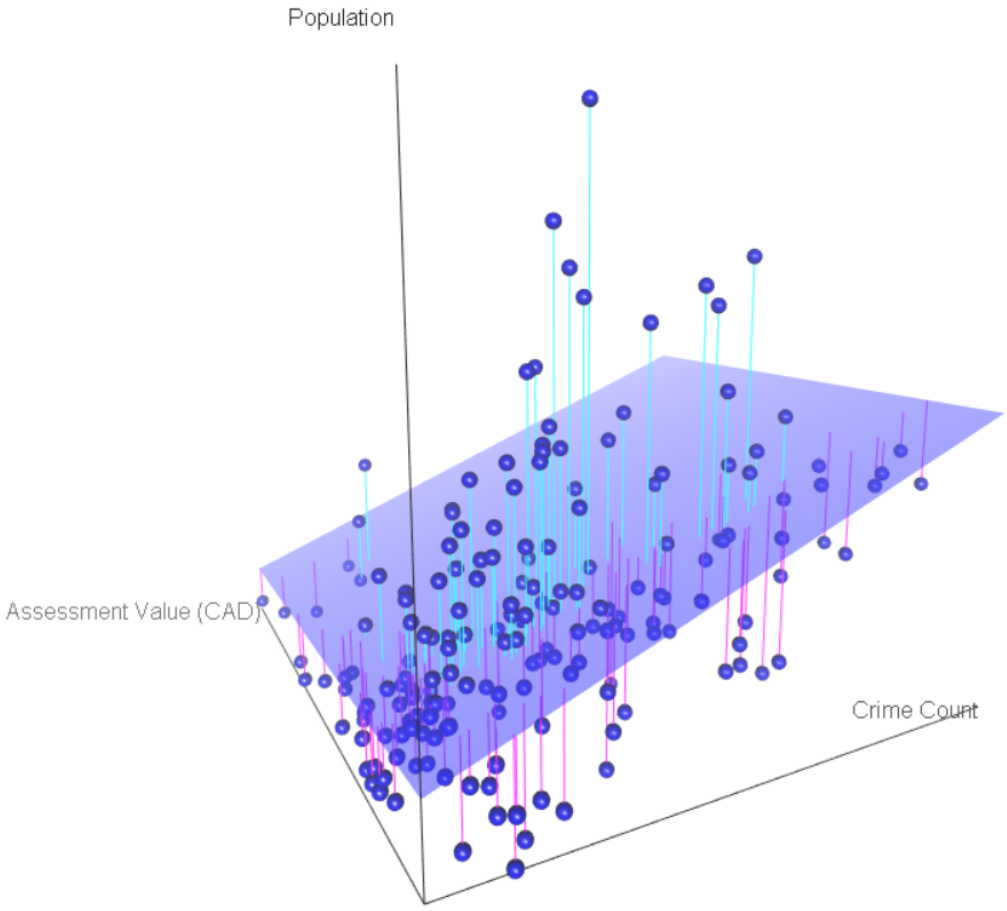
$$Y_{Assess} = A + B_1 X_{Population} + B_2 X_{CrimeCounts}$$

Figure 24: Assessment Values vs Population vs Crime Count 3-D Scatter Plot



As opposed to a straight line fit as shown earlier, since the data here is in three dimensions, a plane is fitted to the data instead. A correlation factor was computed and turned out to have a value of r = *-0.233*.

Figure 25: Assessment Values vs Population vs Crime Count Linear Model



Based on our linear model, the relationship between Assessment Values and Population can be modelled by:

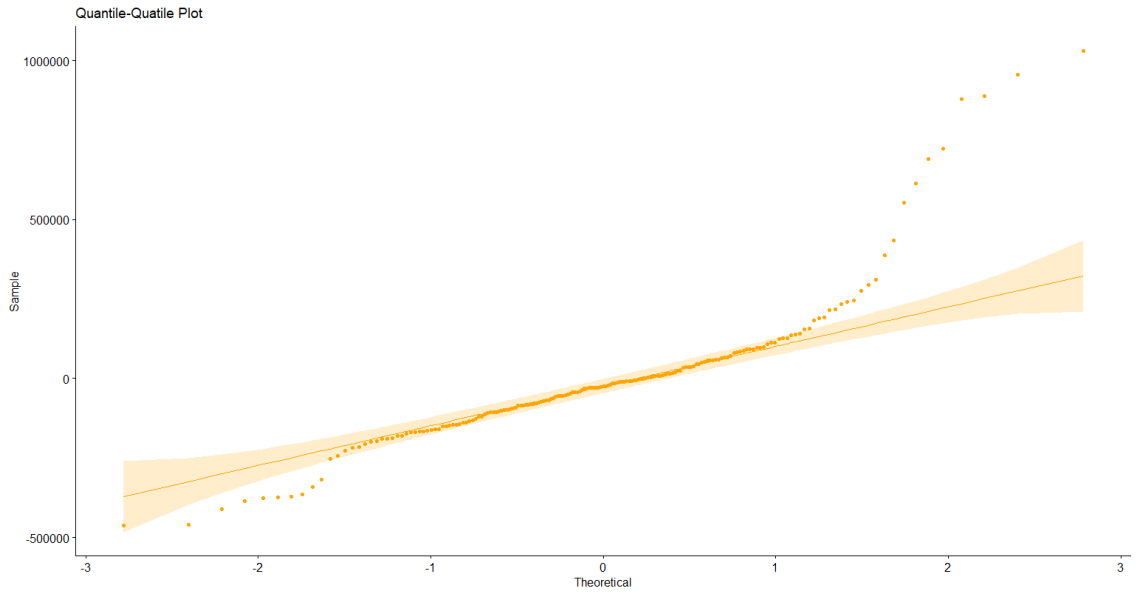
$$Y_{Assess} = 614692.441 + (-2.648)X_{Population} + (-206.943)X_{CrimeCounts}$$

Testing Conditions

Normality of Residuals

For the linear model to be valid, the Assessment Values must be normally distributed. To test this condition, Quantile-Quantile plots were used. This time a QQ plot with confidence bands was used.

Figure 26: Quantile-Quantile Plot.

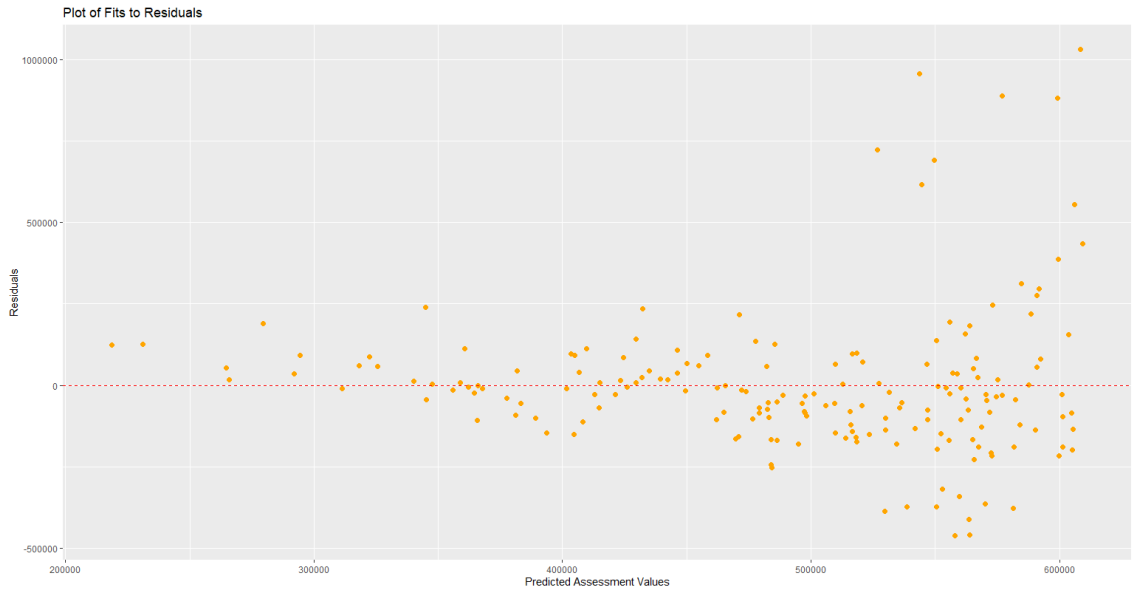


Similar to the earlier situation, most of the data does lie on the straight line, however more of the points are deviating from normality this time. It is still not enough to merit us stopping the analysis here.

Homoscedasticity

The other condition that must hold is the stability of the variance of Assessment Values, otherwise known as the homogeneity of variance. To check for this, we plot the fitted values and residuals on a scatter plot and check the spread.

Figure 27: Plot of Fits to Residual Plot.



Similar to both situations prior, disregarding a few outliers on the top right shows a scatter plot that is bounded above and below by the same Residual value with the residuals scattered in between.

Following that, the coefficient of determination was computed in R, and turned out to have a value of r^2=*0.132*. This value gives the percentage of the variation of the Assessment Values that can be accounted for by their linear dependency on both Population and Crime Counts.

Statistical Significance of the Linear Model

To test for the overall linear appropriateness of the model, the following hypothesis were proposed:

H_0: Assessment Values can not be expressed as a linear function of both Population and Crime Counts ($B_1 = 0$ and $B_2 = 0$)
H_A: Assessment Values can be expressed as a linear function of of both Population and Crime Counts ($B_1 \neq 0$ and $B_2 \neq 0$)

Using R, an F-test was conducted to determine the value of the F-statistic and the p-value. It turned out $F_{1,182,Population} = 9.056$, $F_{1,182,CrimeCounts} = 18.784$ and $p - value_{Population} = 0.00299$,$p - value_{CrimeCount} = 0.0000242$. This is a statistically significant result that allows us to reject the Null Hypothesis based on the fact that both the *p-value* < 0.05.

Based on that, we can say that the Assessment values can indeed be expressed as a linear function of both Population and Crime Count, with a heavier emphasis given to crime counts because of the lower p-value.

As a conclusion for the linear modelling of Assessment Values as a function of Crime Counts and/or Population, it is safe to say that based on our statistical results, it does seem possible to express Assessment Values as a linear function of both the mentioned variables, with the Crime Counts being more substantial in terms of their effect.

The fact that the tail end of the data did not meet Normality conditions can be dismissed based on the Central Limit Theorem. Having around 200 data points is enough to approximate as a Normal distribution.

6 Bivariate Analysis of Population Count and Median Property Assessment

Is there a relationship between Population Count (Variable X) and Property Assessment (Variable Y)?
Lets first look at the scatter plot to get a better idea of our data snapshot by finding the correlation.

(a) Panorama Hills (b) Roxboro

Figure 28: Scatter plot comparison of both communities

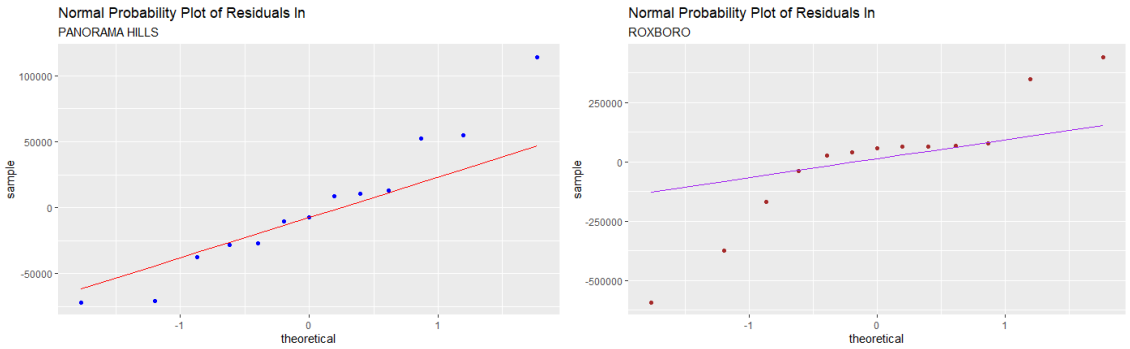
1. Correlation of Population to Property Assessment in PANORAMA HILLS: r = 0.7448511
2. Correlation of Population to Property Assessment in ROXBORO: r =0.2024057

The correlation coefficient for both communities turned out to be positive, with Panorama Hills having a stronger correlation than Roxboro. This positive correlation suggests that as population counts in a community increase, the medial residential property assessments will also increase. We can model this relationship through a linear model as shown below:

Building the Linear Model for Both Communities:

$$Y_{PanoramaHillsAssess} = 248924.2 + 9.109066 \times X_{Population}$$
$$Y_{RoxboroAssess} = 766811 + 1112.444 \times X_{Population}$$

Model Condition 1: The Variable y (Median Property Assessment) must be normally distributed



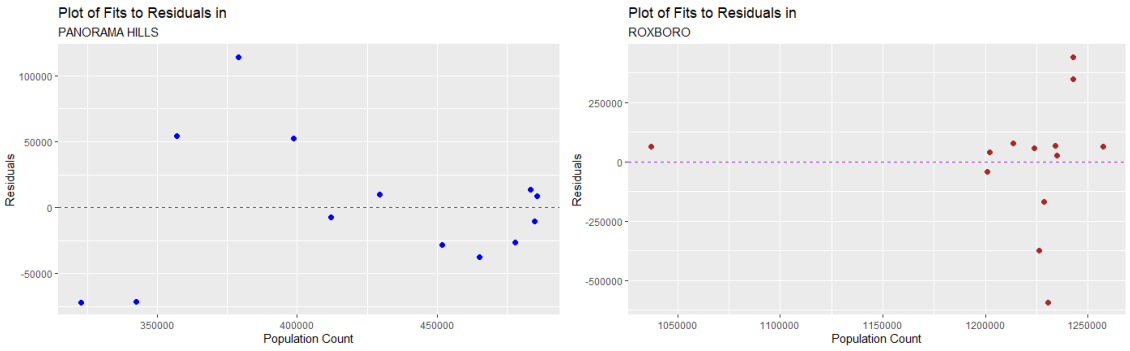
(a) Panorama Hills (b) Roxboro

Figure 29: Normal probability plot of both communities

1) From our normal probability plot in the high population model, we see that the residuals (ei) appear to conform to a normal distribution. There are a few data points towards the end of the lind that separate away, but the vast majority of the points approximate the normal line. As a result, this condition is passed for: PANORAMA HILLS

2) In the normal probability plot of the low population model, the residuals do not appear to conform to a normal distribution. The upper and lower ends vary greatly from the normal line. As a result, this condition is failed to be upheld in: ROXBORO

Model Condition 2: Inspect homoscedascity condition by comparing fitted values with residuals



(a) Panorama Hills (b) Roxboro

Figure 30: Plots of Fitted Values to Residuals

1) There appears to be a wedge shape to the distribution of the points. The distance from the initial points deviate largely from the horizontal line (e = 0), and slowly converge on the centre as the population count increases. As a result, the common variance condition is not fulfilled in the higher populated community of: PANORAMA HILLS

2) This chart also shows a large wedge shape, however it is in the opposite direction when compared to the community with the larger population. As the population count increases, the spread of the residuals also increase. Overall, we see that the homoscedascity condition is not fulfilled in the lower-populated community of: ROXBORO

Confidence Interval Testing for the value of β

Hypothesis Statement for B:

Null Hypothesis H(0): $\beta = 0$ (Community Median Property Assessment cannot be expressed as a linear function of Population)
Alt. Hypothesis H(A): $\beta \neq 0$ (Community Median Property Assessment can be expressed as a linear function of Population)

1 - PANORAMA HILLS

95% confidence interval= $3.694032 \leq \beta \leq 14.5241$

The confidence interval shows the rate at which the median property assessments will change with each increase in community population count. At a 95% certainty, the range of this increase will be between 3.694032 and 14.5241

2 - ROXBORO

95% confidence interval = $-2459.402 \leq \beta \leq 4684.29$

The confidence interval for the lower populated community shows a 95% certainty that the rate at which median property assessments will change is within -2459.402 and 4684.29

Evaluating the significance of confidence intervals through analysis of variance

1 - PANORAMA HILLS

F-Observed = 13.71, P-value = 0.00349

From our analysis of variance, we see that the F-observed value is quite large(13.71). This value represents the ratio between explained and unexplained variance with respect to the fitted values in our linear model. A large F-ratio is a good indicator that the true means of both variables 'population count' and 'median property assessments' are not equal. Our P-value(0.00349) reveals that this is indeed the case and we should reject our null hypothesis in favor of the alternative. This suggests that population count can be expressed as a linear function of median property assessment in PANORAMA HILLS

2 - ROXBORO

$$\text{F-Observed} = 0.47, \text{ P-value} = 0.507$$

In this analysis, the F-value is quite small(0.47). This indicates that the true means of both groups are close to equal. Our P-value(0.507) reveals that we fail to reject our null hypothesis and continue to assume that population count cannot be expressed as a linear function of median property assessment in ROXBORO

Coefficient of Determination

1 - PANORAMA HILLS

$$r^2 = 0.5548032$$

55.48% of all variation in median property assessments can be attributed to its linear dependency of population count

2 - ROXBORO

$$r^2 = 0.04096805$$

4.10% of all variation in median property assessments can be attributed to its linear dependency of population count.

Comparing the two coefficients of determination, we see that communities with larger populations such as Panorama Hills are better at predicting property assessments given a population. While this r^2 value is still moderate at best, it is still a better model than Roxboro.

Using the Models to Predict Future Median Property Assessment Given the Population Projection

Even though both of our models fails to satisfy all the conditions, there are logarithmic transformations that could be used to improve its common variance that is not covered in class. While models may be poor predictors (especially Roxboro), we decided to conduct a predictive test to demonstrate what we have learned in class.

The projected population count for Panorama Hills in 2020 will be 5767.45
The projected population count for Roxboro in 2020 will be 390
These numbers are provided to us through the population projection data in the city of Calgary [1][2]

1 - PANORAMA HILLS

The estimated median property assessment given that the projected population count will be 5767.45 in 2020 is:

$$157992.7 \leq Y_{PropertyAssessment} | x_{Population=5767.45} \leq 44927.8$$

2 - ROXBORO

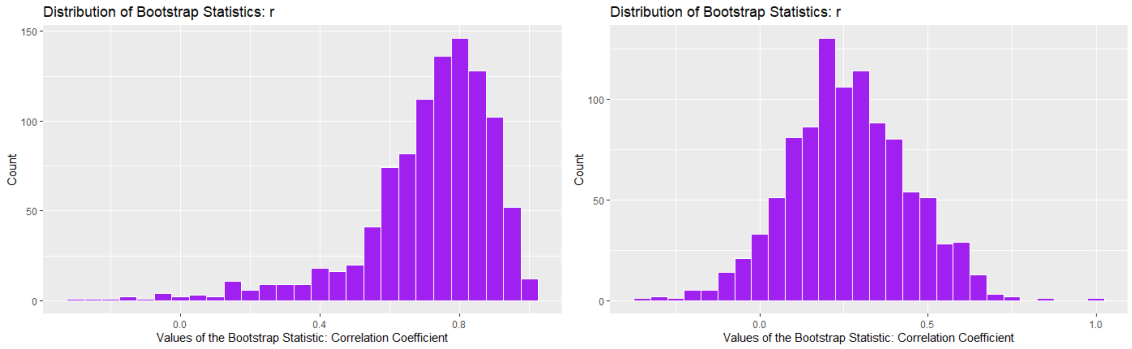
The estimated median property assessment given that the projected population count will be 390 in 2020 is:

$$560568.1 \leq Y_{PropertyAssessment} | x_{Population=390} \leq 1840760$$

Bootstrapping the Means

Since both of our communities did not successfully pass the test for all the conditions (normality & homoscedascity), a bootstrap will be used here to approximate the values of a, b, and r.

Bootstrapping the 95% Confidence Interval of the Correlation Coefficient



(a) Panorama Hills (b) Roxboro

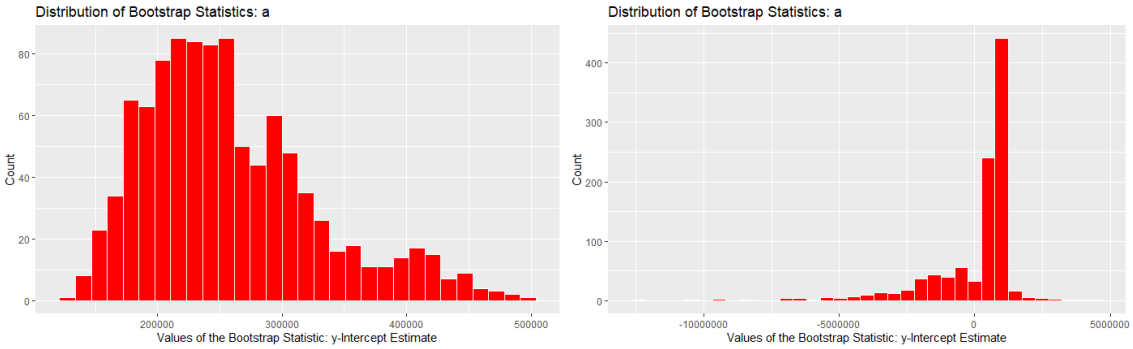
Figure 31: Bootstrap Distribution of r_{boot}

Panorama Hills:
$0.2011 \leq r_{boot} \leq 0.8375$
$\mu\ r_{boot} = 0.7115$

Roxboro:
$-0.0772 \leq r_{boot} \leq 0.6271$
$\mu\ r_{boot} = 0.2645$

The confidence interval represents the mean correlation coefficient between population count and median property assessments for both communities at a 95% certainty. Comparatively, Panorama Hills has a higher correlation coefficient than Roxboro. In practical sense, this suggests that there is a closer relationship between population and property assessment in communities with higher population counts like Panorama, compared to lower populated communities such as Roxboro.

Bootstrapping the 95% Confidence Interval of the y-intercept (a)



(a) Panorama Hills

(b) Roxboro

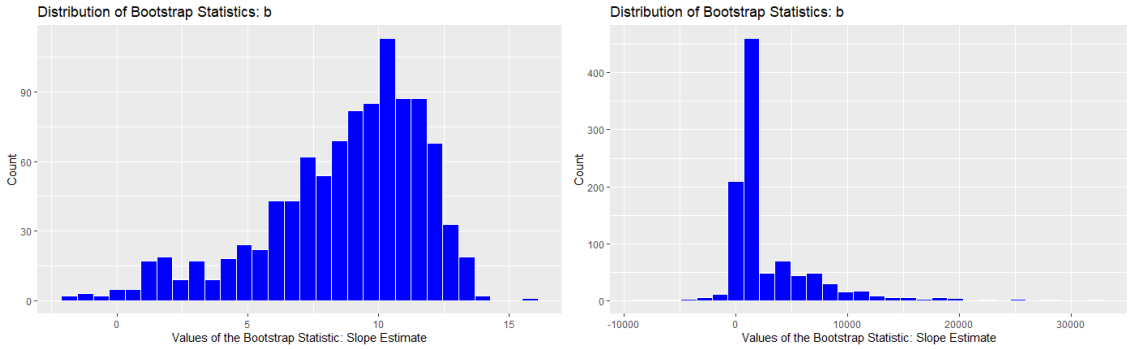
Figure 32: Bootstrap Distribution of a_{boot}

Panorama Hills:
$150478.1 \leq a_{boot} \leq 426353.9$
$\mu\ a_{boot} = 260080.1$

Roxboro:
$-4062839 \leq a_{boot} \leq 1312846$
$\mu\ a_{boot} = 75001.42$

The confidence interval for a_{boot} represents the expected upper and lower value of the median property assessment when population count is zero in each community. We see an overall higher average value for Panorama Hills, and an overall lower for Roxboro. It should be noted that even though the interval for Roxboro drops into the negative, there is no practical meaning having it below zero since property assessments cannot be negative.

Bootstrapping the 95% Confidence Interval of the Slope (b)



(a) Panorama Hills

(b) Roxboro

Figure 33: Bootstrap Distribution of b_{boot}

Panorama Hills:
$1.4181 \leq b_{boot} \leq 13.2026$
$\mu\ b_{boot} = 8.6247$

Roxboro:
$-580.9235 \leq b_{boot} \leq 12675.0474$
$\mu\ b_{boot} = 2765.528$

The b value represents the slope, or the rate at which median property assessments will increase or decrease per increment in population count in each community. We see an overall positive slope in Panorama, suggesting that assessment values will increase with every new resident, whereas the Roxboro captures zero in its interval. This suggests that there is no significant relationship between population and property assessments in lower populated communities such as Roxboro (which is confirmed by our previous case when we looked at the analysis of variance)

Our updated Linear Model based on bootstraped values of A and B:

$$Y_{PanoramaHillsAssessment} = 260080.1 + 8.7166X_{population}$$

$$Y_{RoxboroAssessment} = 63571.42 + 2790.694X_{population}$$

7 Conclusion

This data statistical analysis was an attempt to look into the factors that might affect the property assessment values in the city of Calgary. Trends of assessment values were compared against trends of oil price, population change, and crime counts on a community basis. Our analysis concluded that there is a delayed linear correlation between assessment values and the oil price. We also found that the delay was greater for residential properties than for downtown office building.

Starting with data that does not follow Normality conditions, The Central Limit Theorem was not enough to allow us to assume normality solely based on the size of out data. Therefore, a bootstrap method was adopted to find a trend in the data from 2012 to 2019. It turned out that there was an increase in Assessment Values in that time period, while Population and Crime Counts showed very little to no increase. For all three variables, the standard deviation remained the same from 2012 to 2019.

Regarding the relationship between population count and property assessments, our linear models failed to pass both the normality and homoscedascity conditions. Using the bootstrap approach, these conditions were bypassed and it showed that communities with higher populations (Panorama Hills) had better predictive models than communities with lower populations (Roxboro). Future analysis could include logarithmic transformations to transform the dataset to bypass the common variance condition.

Based on our statistical results, it seems possible to express Assessment Values as a linear function of both the mentioned variables, with the Crime Counts having a more negative effect than Population. The Central Limit Theorem was also envoked to explain away the non-Normality of the tail ends of the data.

8 References

Calgary Open Data. (2018) Calgary's Population 1958 - 2019. Available from: https://data.calgary.ca/Demographics/Civic-Census-Results-1958-2019/rmai-qvzh [Accessed 27th September 2019]

Calgary Open Data. (2018) Property Assessments. Available from: https://data.calgary.ca/dataset/Property-Assessments/6zp6-pxei [Accessed 27th September 2019]

Calgary Open Data. (2018) Community Crime and Disorder Statistics. Available from: https://data.calgary.ca/Health-and-Safety/Community-Crime-and-Disorder-Statistics/848s-4m4z [Accessed 27th September 2019]

Campbell, Douglas A. "Crime and Property Values." The Free Library, University of Memphis, 2007, retrieved September 27, 2019 from https://www.thefreelibrary.com/Crime%20and%20property%20values.-a0170114047

Glink, I. (2013, May 20). Can a rise in crime increase your property value? Retrieved September 27, 2019, from https://www.cbsnews.com/news/can-a-rise-in-crime-increase-your-property-value/

Investing.com. (2019) Crude Oil WTI Futures Historical Data. Available from: https://ca.investing.com/commodities/crude-oil-historical-data [Accessed 28th September 2019]

Maximino, M. (2017, February 16). The impact of crime on property values. Retrieved September 27, 2019, from https://journalistsresource.org/studies/economics/real-estate/the-impact-of-crime-on-property-val
ues-research-roundup/

Property Assessments 2019, City of Calgary, viewed September 30, 2019, <https://www.calgary.ca/PDA/Assessment/Pages/Property-Assessment.aspx>