Are Older Condos in Toronto Cheaper?

Hassan Shah
ECO375: Empirical Project

Introduction

As inhabitants of one of the most expensive cities in the world, Torontonians often wonder if they will ever be able to own a home. There are plenty of mortgage options in the market available to Canadians. However, with a growing population, rising interest rates, and an ongoing cost of living crisis, anxiety surrounding the desire for home ownership ensues. This paper will attempt to remove some of those anxieties by exploring the characteristics that affect the price of Toronto condos. More specifically, this paper will explore whether the cost of condos in Toronto falls as they age, with a focus on amenities.

Some intuition regarding the variability in the price of housing options due to certain characteristics is often valid. For example, one would expect a larger condo in terms of footage to be priced higher than a smaller one. This expectation almost always holds true for any kind of housing option. Yet, the effect of a characteristic like age is ambiguous. Do older homes cost more because maintenance fees rise, or does a high demand for newer condos lead to the older condos becoming more affordable? That is the question this paper will attempt to answer.

The Data

The study will utilize a cross-sectional dataset consisting of 17 different variables pertaining to the material and non-material features of 315 condos in downtown Toronto. These include transaction price, asking price, age of condos, number of bedrooms and bathrooms, amenities such as gyms and movie rooms, etc. The variables in focus will include three non-material features: price, maintenance (fee) and age; material features will include 8 amenities: bedrooms, bathrooms, pool, hottub, gym, movieroom, pet and parking.

I found some interesting trends in the provided data set. Firstly, there were instances where regressors had the exact same value as the total sampling size. Pet, for example, a binary variable representing whether the condo building allowed pets, occurred in every sample unit. This would induce perfect collinearity (violation of MLR3); thus, the variable was excluded. Furthermore, some regressors had a strong correlation with one another. Condo units in the sample that had a pool invariably had a hot tub facility, prompting the exclusion of one variable due to concerns of multicollinearity.

As most of the variables included in the regression were either binary variables or provided in non-numeric form, I transformed them into a numeric form prior to analyzing the model. Variables with the number 1 at the end, such as age1, pool1, pet1, etc, have been transformed from their character forms to numeric forms. Additionally, I created multiple other transformations, some of which were crucial to add to the depth of the analysis, while others aided in verifying the assumptions of the multiple linear regression model. For the purpose of introducing depth to the analysis, I created an age-squared variable (age1_sq). Since the age variable forms a crucial focus of this paper, I was interested in analyzing it in more than one way. The significance of the variable will be discussed in the Discussion section. Further, I created logged variables for my regressors to prove MLR5; however, these were eventually replaced by an alternative technique that used robust standard errors for a more effective solution.

	Min	Max	Mean	Std. Dev
price	353000	2925000	758797.0127	255565.1996
age1	1	32	11.95238	8.43678
age1_sq	1	1024	213.8127	290.23285
maintenance	238	3371	627.75238	307.31048

Table 1: Summary Statistics of non-binary regressors

The table above depicts useful summary statistics of the non-binary variables in my final model. In addition to these, I included certain binary variables: den1 (n = 153/315), pool1 (n = 216/315), parking1 (n = 169/315) and finally, bathrooms1.

Every condo had a bathroom; however, the breakdown was the following: 1 bathroom (n = 191/315), 2 bathrooms (n = 123/315), and 3 bathrooms (n = 1/315).

Methodology and Analysis

The objective of the current study was to create a linear regression model based on relevant amenities and age related regressors to forecast condo prices. This was achieved by constructing models with all suitable variables that produced significant results without violating any MLR assumptions. For the purpose of this study, I chose to use the Ordinary Least Squares method of estimation. Initially, a preliminary model (mod0) was constructed that regressed price on 11 different independent variables.

$$\text{price} = \beta_0 + \beta_1\,\text{footage1} + \beta_2\,\text{maintenance} + \beta_3\,\text{den1} + \beta_4\,\text{bathrooms1} + \beta_5\,\text{hottub1} + \beta_6\,\text{bedrooms1} + \beta_7\,\text{parking1} + \beta_8\,\text{age1} + \beta_9\,\text{pool1} + \beta_{10}\,\text{pet1} + \beta_{11}\,\text{movieroom}\ \textbf{(mod0)}$$

As an example, the Bedrooms1 variable was part of the model at this stage; however, it was eliminated as it did not generate a significant result. In order to produce an unbiased model, additional non-significant variables were progressively eliminated from the model. For instance, the variable pet1 had no variation, which violated MLR3. Thus, pet1 had to be eliminated.

	Coefficient	Std. Error	T-stat	Prob
Intercept	325638.83	76773.55	4.242	2.94e-05 ***
footage1	499.87	55.01	9.087	< 2e-16 ***
maintenance	238.45	39.2	6.082	3.53e-09 ***
den1	8816.23	13741.12	0.642	0.521617
bathrooms1	77263.76	19947.83	3.873	0.000131 ***
hottub1	-58330.69	17574.61	-3.319	0.001012 **
bedrooms1	NA	NA	NA	NA
parking1	71924.93	16365.56	4.395	1.53e-05 ***
age1	-14875.82	1030.25	-14.439	< 2e-16 ***
pool1	NA	NA	NA	NA
pet1	NA	NA	NA	NA
movieroom1	-3190.59	70170.12	-0.045	0.963763
Multiple R-squared	0.7979	F-Stat	151.1	
Adjusted R-squared	0.7927	P-value	<2.2e-16	

Table 2: Regression output for preliminary model (mod0)

At this point, the den1 transformed variable did not produce significant results either. Yet, since dens are generally considered an amenity, it was retained in the model. Instead, I experimented further by excluding the footage1 variable and re-running the regression to find out that the statistical significance of den1 became important. A justification is provided in the discussion section for this decision. Finally, to verify if the effect of age is in the same direction in both the restricted and unrestricted regressions, I regress age on price.

$$\text{price} = \beta_0 + \beta_1 \text{ age1 } \textbf{(mod1)}$$

	Coefficient	Std. Error	T-stat	Prob
Intercept	815599	24726	32.98	< 2e-16 ***
age1	-4752	1691	-2.81	0.00526 **
Multiple R-squared	0.02461	F-Stat	7.898	
Adjusted R-squared	0.0215	P-value	0.00526	

Table 3: Restricted regression output (mod1)

This proved that the isolated and combined effect of age is in the same direction. Thus, the second rendition of the model was tested, yielding surprising results. The hottub1, gym1, and movieroom1 variables had lost significance. I suspected this was due to issues of multicollinearity, so I checked for patterns in the data set. Unsurprisingly, all condos that had a pool also had a hot tub, but I was unsure as to what variables gym1 and movieroom1 were correlated with.

	Coefficient	Std. Error	T-stat	Prob
Intercept	401158.77	86102.97	4.659	4.75e-06 ***
maintenance	473.35	33.78	14.013	< 2e-16 ***
den1	37957.42	15069.39	2.519	0.0123 *
bathrooms1	149892.24	20582.27	7.283	2.79e-12 ***
hottub1	NA	NA	NA	NA
parking1	91845.06	18416.58	4.987	1.03e-06 ***
age1	-15870.63	1166.45	-13.606	< 2e-16 ***
pool1	-91239	19484.31	-4.683	4.26e-06 ***
gym1	-29327.85	61835.79	-0.474	0.6356
movieroom1	64860.5	96784.7	0.67	0.5033
Multiple R-squared	0.7436	F-Stat	110.9	
Adjusted R-squared	0.7369	P-value	< 2.2e-16	

Table 4: Unrestricted regression output (mod2)

So, I found it imperative to test the correlations between my regressors such that I would not have to discard all of the variables relating to amenities, including pool1, that I retained from mod0 for further testing. Subsequently, I ran a correlation analysis and printed a correlation matrix. As I suspected, pool1 and hottub1 had a high correlation with each other, but so did movieroom1 and gym1. Between pool1 and hottub1, I retained pool1 after assigning a value of 1 and 0 to both variables and using Excel’s random number generator feature to decide. The gym1 and movieroom1 variables had to be taken out immediately as a result of high covariance as well as low variance.

	Coefficient	Std. Error	T-stat	Prob
Intercept	437591.66	33391.59	13.105	< 2e-16 ***
maintenance	470.64	33.35	14.114	< 2e-16 ***
den1	38506.27	14928.27	2.579	0.0104 *
bathrooms1	150439.51	20513.07	7.334	1.99e-12 ***
parking1	92591.64	18181.47	5.093	6.16e-07 ***
age1	-15874.91	1139.72	-13.929	< 2e-16 ***
pool1	-91893.3	19271.2	-4.768	2.87e-06 ***
Multiple R-squared	0.7432	F-Stat	148.6	
Adjusted R-squared	0.7382	P-value	< 2.2e-16	

Table 5: Unrestricted regression (mod2)

After excluding all problematic variables, I finally acquired a working model that is significant in all regressors. Yet, it did not represent the effect of age on price in the depth that I desired to study. For a deeper analysis, I added an age1 squared variable to the existing model and added ran one last regression that I believed would suffice to accurately predict the price of Toronto condos.

	Coefficient	Std. Error	T-stat	Prob
Intercept	492028.79	38126.08	12.905	< 2e-16 ***
maintenance	477.15	33.05	14.439	< 2e-16 ***
den1	28819.89	15143.39	1.903	0.057958 .
bathrooms1	138884.04	20679.54	6.716	9.05e-11 ***
parking1	94802.06	18181.47	5.269	2.58e-07 ***
age1_sq	307.17	17990.89	2.854	0.004614 **
age1	-25644.76	3604.14	-7.115	7.91e-12 ***
pool1	-74056.34	20050.61	-3.693	0.000262 ***
Multiple R-squared	0.7499	F-Stat	131.5	
Adjusted R-squared	0.7442	P-value	< 2.2e-16	

Table 6: Final unrestricted regression model (mod3)

Within this improved model, the enhanced impact of age along with other amenities on property price was examined. Unfortunately, den1 loses significance at the 5% level, however I will still keep it in the model, considering it measures the effect of the den amenity as well as ‘extra footage,’ as discussed in the discussion section.

Following the inclusion of a directly transformed variable, such as age1 squared, there is the caveat of multicollinearity between the variable and its basic form (age1). I tested this by obtaining the VIF scores of my regressors, which are mentioned below.

maintenance	den1	bathrooms1	parking1	age1_sq	age1	pool1
1.937868	1.079853	1.980739	1.672208	18.338565	17.374337	1.633262

Table 7: VIF scores

Conventionally, VIF scores of > 5 are considered optimal. For all regressors other than age1 and age1 squared, VIF scores are highly optimal. Since the age1 squared variables just takes all the values from the age1 column and squares them, this test detects high multicollinearity. I do not consider this as inherently problematic to my model; I further elaborate on this in the discussion section.

As I was satisfied with my models and considered them complete, I moved on to testing for joint significance of the restricted and unrestricted models. This was facilitated by the F-test feature in RStudio. Using the ANOVA function, I performed an F-test on mod3 and mod1. Below are the results.

F-test (mod3, mod1):

- F value = 148.35
- P value = 2.26e-16

Finally, I moved on to testing the **MLR assumptions**, which would verify the validity and reliability of the model. MLR1, pertaining to the linearity of the model, was attained through obtaining significant regressors in both the final restricted and unrestricted regressions (mod1 and mod3). MLR2 could not be verified empirically as I did not have control over the data collection process, nor was it stated that the sampling was done randomly. Hence, it will be assumed that MLR2 holds. MLR3, requiring there to be no perfect multicollinearity between regressors, was tested earlier when the VIF scores of mod3 were calculated. The issue of high VIF between age1 and age1 squared will be overlooked for the purposes of this study; further elaborated in the discussion section.

MLR4 states the expected value of the error term conditional on x should be equal to zero. To verify MLR4, I extracted the residuals and fitted values from my unrestricted model to create a residual plot. The line of best fit, as seen in figure 1, appears to be centered exactly at zero, or very close to it. Thus, this assumption also holds.

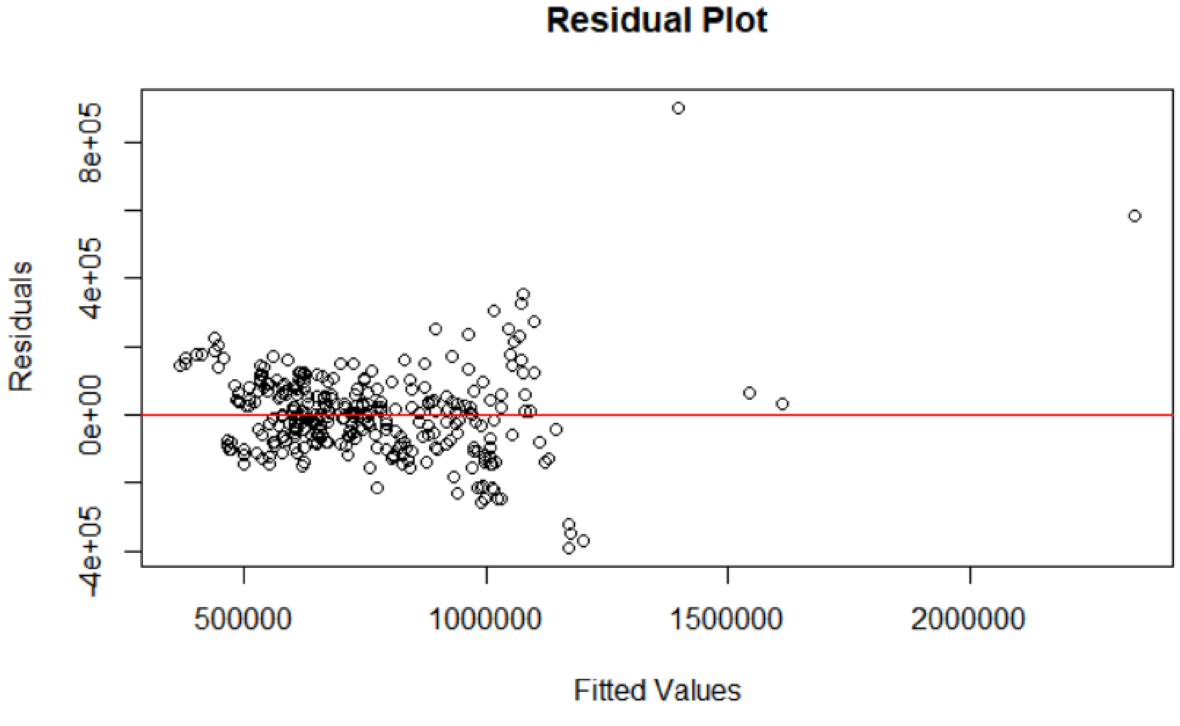

Figure 1: Residual Plot

MLR5 was tested using the Breusch-Pagan test. MLR5 carries great significance. It requires that the errors in the model are dispersed in a homoscedastic fashion. If this is violated, then under the Gauss-Markov theorem, OLS is not the most efficient estimator. The studentized Breusch-Pagan test suggested:

- BP = 75.578
- P value = 1.094e-13

This is problematic for the model's efficiency as MLR5 is violated. A p-value close to zero leads to the rejection of the null hypothesis of homoscedasticity. In order to correct this, I first analyzed the direction of the skew using a Q-Q plot for clarification purposes.

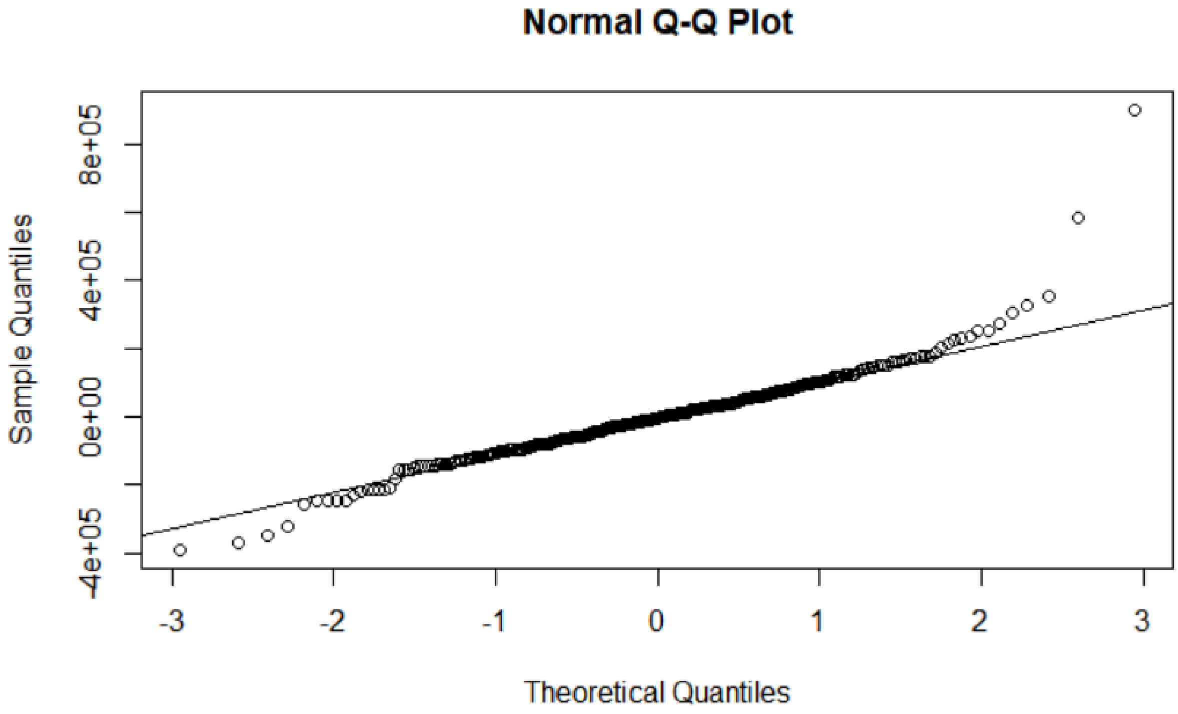

Figure 2: Q-Q Plot showing direction of skew of errors

The plot provided visual proof of the distribution of residuals not being normal, whereby the errors had a left skew. As part of my initial correction effort, I log transformed all my regressors from mod3, then I ran the regression. Surely enough, the values of my logged variables fell to the negative infinity range, because the majority of my variables (for amenities) are dummies, taking a value of 0 or 1.

Proceeding with caution, I only used the log transformations for maintenance, age1 and age1 squared variables this time around. The model appeared to be significant, but it failed the test once again, with its residuals still skewed.

Another fix for heteroscedasticity that I attempted is using robust standard errors. As shown in Table 8, the coefficients of this new regression (mod3_robust) are almost exactly identical to those in the previous final regression model (mod3). With updated standard errors, I can finally accept that MLR5 has been met.

	Coefficient	Std. Error	T-stat	Prob
Intercept	492028.792	48883.711	10.0653	< 2.2e-16 ***
maintenance	477.15	93.615	5.0969	6.047e-07 ***
den1	28819.889	17955.388	1.6051	0.109504
bathrooms1	138884.04	23385.939	5.9388	7.768e-09 ***
parking1	94802.055	29070.269	3.2611	0.001235 **
age1_sq	307.17	152.087	2.0197	0.044283 *
age1	-25644.762	5752.514	-4.458	1.161e-05 ***
pool1	-74056.342	15425.054	-4.801	2.470e-06 ***

Table 8: Regression with updated Robust Standard Errors

Moving on to MLR6, which argues the normality of the error term. For this, I created a density plot of the model's residuals. The figure is posted below.

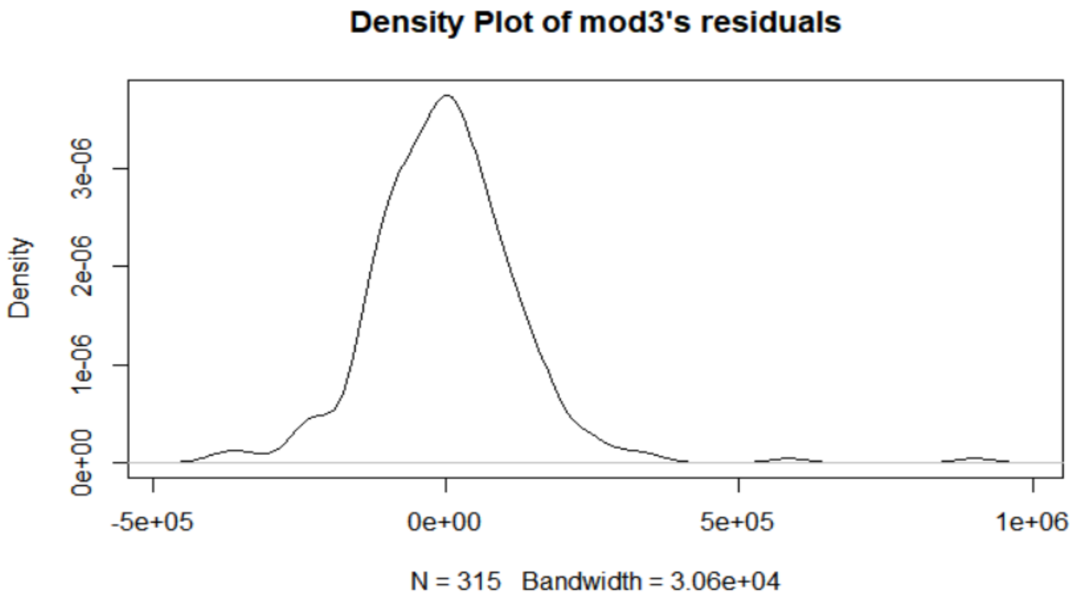

Figure 3: Density Plot

With an almost normal distribution, centered close to zero, MLR6 can be argued to hold as well.

Discussion and Conclusion

First, I will discuss why I took certain steps that would not be considered conventional but were crucial to the final result of my analysis.

Inclusion of the pool1 variable:

When I performed a preliminary regression, as depicted in table 2, pool1 did not yield a coefficient, suggesting that the presence of a pool had no effect on condo prices. Yet, I included the variable in further analysis (mod2), as seen in table 4. This was not a mistake as I intentionally added pool1 to verify the results from the earlier regression. The first time I questioned the result (from mod0) was when I saw a negative coefficient for hottub1, suggesting that the presence of a hot tub decreases the prices of condos. This was highly counter intuitive to me. Further, when I noticed the pattern of pools and hot tubs coexisting in the same condos in the sample, I was further skeptical about retaining hottubs1. Thus, I replaced hottub1 with pool1, which became a significant part of my regression model through to the end.

Replacement of footage1 with den1:

In the preliminary regression (mod0), den1 has no significance, yet it was retained in lieu of footage1, which had a high significance of three stars as seen in table 2. In my defense, I replaced the variables because I interpreted the den variable as an extra room which adds footage to a condo unit. A limitation of the study is that I could not control for footage, which would have made the interpretation of den1 easier to comprehend, as I have interpreted it, yet units that have a den could be thought of as having more space (footage). Furthermore, den1 was an amenity variable, which is what the focus of the paper is on. If I kept footage, the significance of den1 would shrink to nothing. Therefore, I decided to remove footage1 and replace it with den1.

High multicollinearity between age1 and age1 squared:

As seen by the VIF scores in table 7, age1 and age1 squared have an extent of multicollinearity. However, I proceeded with my analysis regardless of this violation of MLR3. The justification I provide for this relates to the nature of the variables. I will agree that when variables like, for example, French fries and Poutine have a high correlation, that certainly indicates multicollinearity of a problematic nature. On the other hand, when a variable and its transformed version have high correlation, it doesn't affect the model foundationally as their correlation will be high by construction. The benefits that the age1 squared variable offers (as discussed below) outweigh its costs (high VIF), so I argue that this does not become problematic for the model in itself.

Inclusion of den1 despite losing significance:

In table 8, showcasing the robust standard errors, the den1 variable loses significance at the 10% level. Most econometric models would discard such a variable. Yet, I retain den1 because of its importance to the model. Since the beginning of my analysis, I have preserved den1 as an amenity variable encompassing the effect of added footage to the condo alongside having an extra den. Without having to use robust standard errors, I found den1 significant at the 10% level. I believe this variable is a special case that deserves an exception; thus, I have kept it till the end.

Importance of age1 squared:

I mention in the data section that I desire to measure an enhanced effect of age on condo prices. The age1_sq variable facilitates that desire by giving me an outlook on the relationship between age and price through a different lens. As seen in table 6, the age1 variable has a negative relationship with price. It could be argued that people have a preference for newer houses, or that maintenance costs of older houses go up such that the price of condos is negatively related to their age. Inversely, the age1_sq variable establishes a relatively weaker but positive link between age and price. This could be interpreted as a 'vintage effect,' whereby people prefer older housing, perhaps because older condos are larger or that they have some special architectural features that newer housing options don't showcase.

Validity, Consistency, and Asymptomatic Normality:

Considering that MLR1-4 hold, even in the face of these blaring exceptions mentioned above, I can state with confidence that the estimators are efficient measures of the effect of independent variables on the price of Toronto condos. Referencing the density plot in figure 3, I can also claim that if the sample size were to increase to infinity, my estimators would produce results consistent with the population statistics and asymptomatic normality would be achieved.

Given these limitations, it is imperative to mention that the variables I have worked with in this study are not the only factors that determine the price of condos. Most especially, interest rates in Canada are a large determining factor of condo prices as they dictate mortgage policies. If the inflation and interest rate variables could be controlled for, that would produce even more accurate estimates that would produce highly accurate results.