Analyzing the Determinants of Condo Unit Prices in Downtown Toronto

UMER ABBASI

April 06, 2023

Table of Contents

Analyzing the Determinants of Condo Unit Prices in Downtown Toronto

Umer Abbasi

Introduction

The objective of this study is to comprehensively investigate the various factors that determine the price of a condo unit in Downtown Toronto. While square footage, number of bedrooms, and bathrooms are commonly believed to be the primary determinants of condo prices, Toronto's dynamic and competitive housing market suggests that there may be other influential factors at play. This study aims to provide valuable insights to help individuals better understand the significant features that can impact a condo unit's price, whether positively or negatively. This report will start by providing an overview of the data and detailing the methodology employed for the empirical analysis. Afterwards, the obtained results will be presented, followed by a brief discussion of the findings and limitations of this study.

When searching for related studies, a study by Marcus Allen [1] came up. In that study, the author identified several significant factors that positively influence the price of a condo unit in Florida, including the square footage of the livable area, the presence of a pool or tennis facility, available parking, whether the condo is situated within a gated community, and the availability of nighttime security [1]. Another study by Kim et al. [2] determined having a view of the sea, a rooftop terrace, and nearby access to a Mass Transit Railway (MTR) in Hong Kong have a positive effect on condominium prices [1]. However, they determined having a balcony, a swimming pool, and a shopping plaza adjacent to the condo to be factors that hurt the price of a condominium unit [1]. When comparing the findings of these two studies, it becomes apparent that the determinants affecting a condo unit's price vary and, more so, seem to differ by location. For instance, while a swimming pool facility adds value to a condo unit in Florida [1], it appears to lower the property's value in Hong Kong [2]. Therefore, it is imperative to acknowledge that our present study, which investigates the determinants of condo prices in the downtown area of Toronto, may yield contrasting results.

Methods

The data used in the study consists of information from 315 downtown Toronto condominium units from October 01, 2021, to September 30, 2022, and includes variables such as price, building ID, maintenance, asking price, parking, bathrooms, floor, age, pool, hot tub, gym, movie room, pet, studio, den, bedrooms, and square footage. We used multiple linear regressions using the ordinary least squares (OLS) method to investigate the research question, as this technique allows us to establish a relationship between the price of a condo unit (response variable) and its corresponding features (predictor variables). We would like to come up with a linear model that could best explain the results. We used R to load the dataset of condominium prices and their characteristics and performed exploratory data analysis (EDA) to identify any potential issues that may require addressing later in the analysis. Next, we split the data into two approximately equal parts, with one half designated as the training set and the other as the test set. The test set will be used to validate our model and assess its generalizability. To ensure that the training and test sets are comparable, we examine their means and standard deviations and determine they are not too different from each other, as such differences could indicate potential issues in our analysis.

Then, we created a full linear model using the training set, with price as the response variable and the other factors as predictor variables. We then assessed the model assumptions, MLR 1-6, using three diagnostic plots: residual vs. fitted plot, residual vs. predictor plot, and a normal quantile-quantile (q-q) plot. MLR 1 and 2 will hold if the data is linear in its parameters and comes from a random sample. MLR 3 will hold if there is no perfect multicollinearity in the data, which will be determined through variance inflation factor (VIF) values. Predictors with VIF > 5 will be considered problematic. MLR 4 will be verified using the fitted vs. residual plot. MLR 5, homoskedasticity assumption, will hold as we will be calculating robust standard errors, which are robust to heteroskedasticity, for our linear models. MLR 6 will hold if the scatter plots on the q-q plot roughly follow the diagonal q-q line. Model assumption violations would be corrected using a transformation, using the natural-log transformation where appropriate. Any transformations that have been applied will also be applied to the test set.

Although all assumptions are expected to hold, MLR 3 may be an exception. Ideally, we aim to use a model with no, or minimal, multicollinearity violations. If MLR 3 is violated (VIF > 5), we will address this issue by iteratively removing predictors with the highest VIF values and rechecking the VIF. This process will be repeated until a model with VIF < 5 is obtained. Note if the variables being removed carries significance, or are essential to this study, then the variable would be kept and discussed under the limitations in this study.

When all the model assumptions, MLR 1-6, hold we will use the results of a t-test on the full model that will help us with the creation and selection of the best linear model that could be used to answer this study. We will create three models in total: a full model and two other models by removing insignificant variables from the full model and generating statistical summaries for each of them. We will compare the R^2 value, and adjusted coefficient of determination (Adj. R^2) value to determine the best model. After an appropriate linear model is determined, this model will be validated with its respective linear model from the test data set. We expect the coefficients from the training set to fall within 2 standard deviations of the coefficients from the test set, and we aim to see the same significant predictors in both test and training sets for each model. Once the model is validated, the results could be made generalizable to the population outside of the sample. We conclude with the validated model as the appropriate linear model to answer the study. The significant parameters in this model as determined by the t-test serve as the determinants that influence condo prices in Downtown Toronto.

Results

Through an EDA of our data on 315 condos in Downtown Toronto from October 01, 2021, to September 30, 2022, the histograms as shown in Figure 1 reveals price, maintenance, floor, studio, and footage to be slightly right skewed. Whereas, pool, hot tub, gym, and movie room, as categorical variables, seem left skewed. The skews may impact our results if they are deemed problematic when checking our model assumptions.

Figure 1a – Histogram displaying the distribution of each data in the original dataset

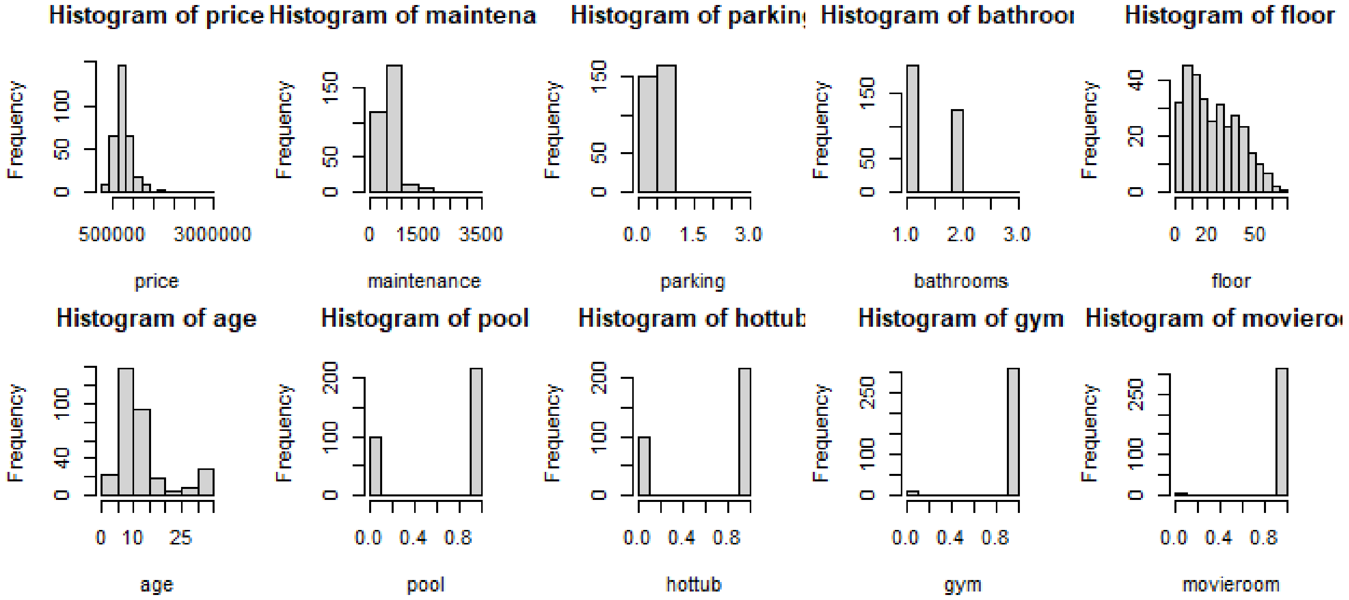
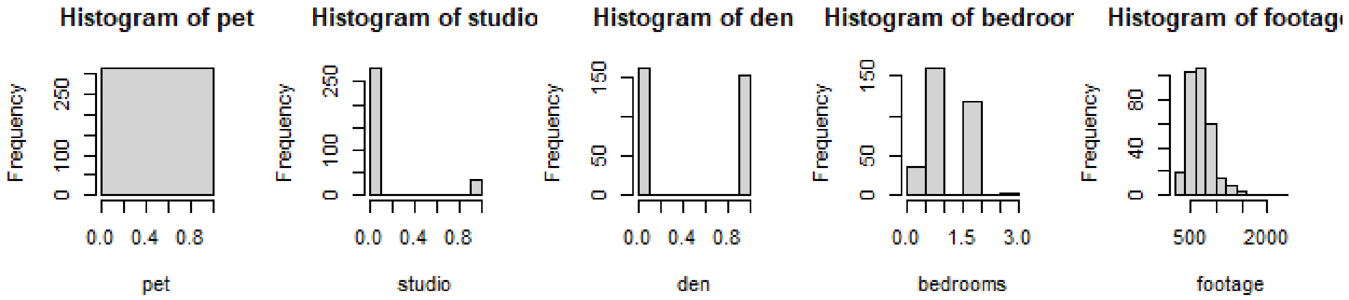


Figure 1b: continued



The boxplots in Figure 2 illustrates several outliers in our data, with price, footage, and maintenance being the most prominent.

Figure 2 – Boxplot displaying variability and outliers for select variables in the original dataset

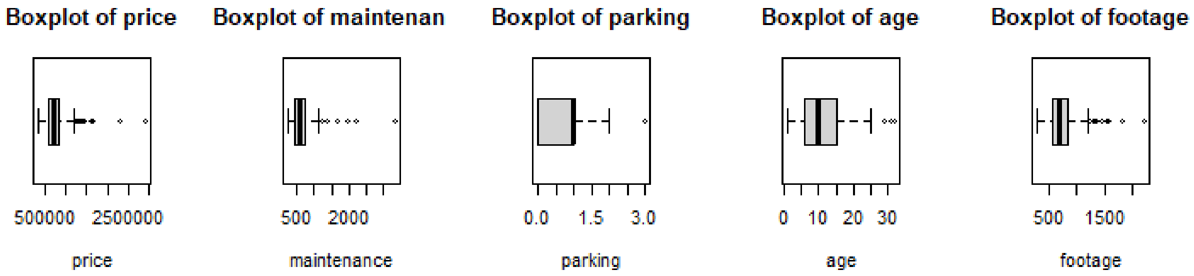


Table 1 presents the summary statistics of our data, which was divided into two roughly equal sized groups - a training dataset with 157 observations and a test dataset with 158 observations. Our analysis indicates that the two datasets are comparable, with no significant differences between them, ensuring that our model will be verified correctly without being affected by the split in data in the datasets.

Table 1 - Summary statistics in training and test dataset, where training set has size 157 and test set 158, approx. 50/50 split.

Variable	mean (s.d.) in training	mean (s.d.) in test
price ($)	792676.6 (297938.26)	725131.85 (200342.06)
maintenance	640.357 (352.421)	615.228 (255.251)
parking	0.554 (0.548)	0.519 (0.501)
bathrooms	1.408 (0.506)	1.386 (0.488)
floor	25.955 (16.518)	22.481 (14.496)
age	11.299 (7.861)	12.601 (8.95)
pool	0.675 (0.47)	0.696 (0.461)
hot tub	0.675 (0.47)	0.696 (0.461)

Variable	mean (s.d.) in training	mean (s.d.) in test
gym	0.981 (0.137)	0.968 (0.176)
movie room	1 (0)	0.981 (0.137)
pet	1 (0)	1 (0)
studio	0.089 (0.286)	0.133 (0.341)
den	0.516 (0.501)	0.456 (0.5)
bedrooms	1.306 (0.647)	1.253 (0.686)
footage	727.726 (270.744)	687.323 (205.755)

Full Economic Model: Price + Maintenance + Parking + Bathrooms + Floor + Age + Pool + Hot Tub + Gym + Movie Room + Pet + Studio + Den + Bedrooms + Footage

After verifying the MLR conditions and VIF, a full linear model was created based on the economic model. However, the residual vs. fitted plot in Figure 3 showed a potential issue with homoskedasticity (constant variance), which was corrected by applying a log transformation to the price variable. The residual vs. predictor plots and normal q-q plots showed no other condition violations or issues with normality. Correlations between variables revealed issues with the pet, movie room, and hot tub variables. Pet and movie room were identified as problematic due to their constant values, which could also be observed in Table 1. Additionally, hot tub was problematic due to having identical values to the pool variable, which induce perfect collinearity. To address these issues, the three variables were removed. Calculating the VIF revealed multicollinearity violations with five variables, but after removing the bedrooms and bathrooms, which had the largest VIF values, the issues were resolved. Since all model conditions were met, this resulted in a candidate model, named Model 1, to answer the research question. Appendix 1 provides further details on the identified VIF violations.

Figure 3a - Residual plots and normal q-q plot for the full linear model to verify model assumptions

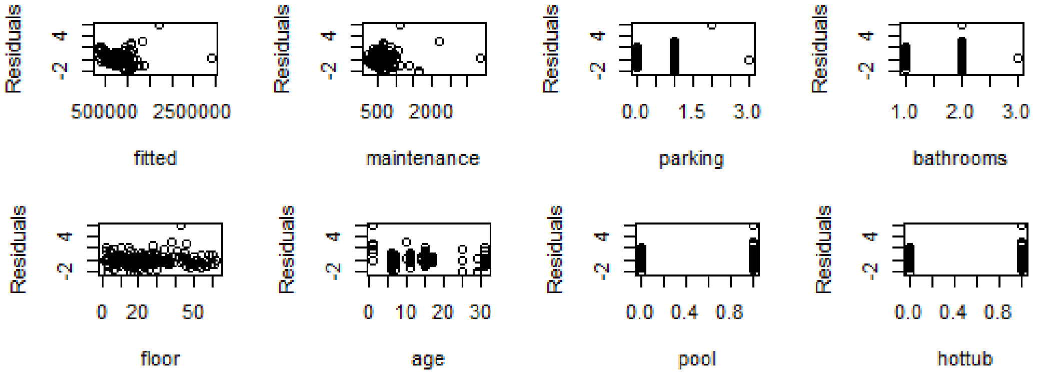
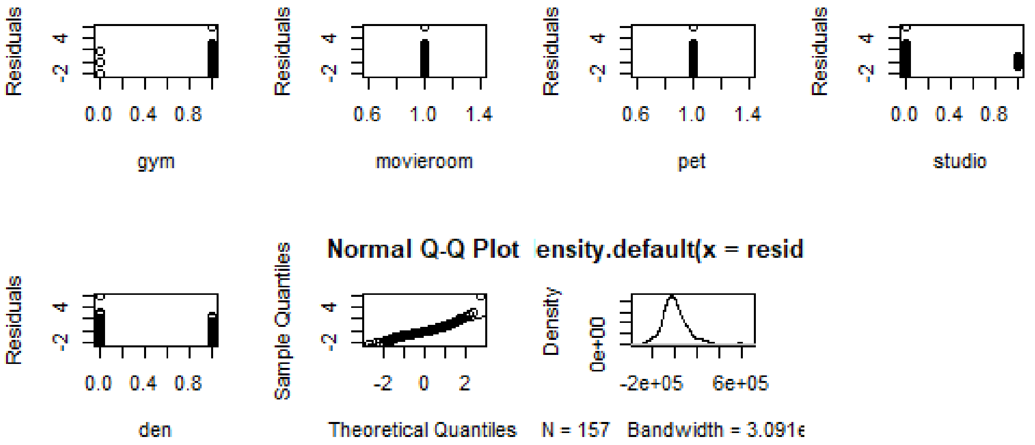


Figure 3b: continued



Two additional linear models were created. One by removing two of the variables with the largest p-value, den, and maintenance, referred to as Model 2, and the other by removing all insignificant variables at the 10% significance level from the full model – Model 1, as Model 3.

- Table 2 Summary of candidate models

Candidate Model	Regression Model
Model 1	= $\beta_0 + \beta_1$Maintenance + β_2Parking + β_3Floor + β_4Age + β_5Pool + β_6Gym + β_7Studio + β_8Den + β_9Footage + u
Model 2	= $\beta_0 + \beta_1$Parking + β_2Floor + β_3Age + β_4Pool + β_5Gym + β_6Studio + β_7Footage + u
Model 3	= $\beta_0 + \beta_1$Parking + β_2 Floor + β_3Age + β_4Pool + β_5Studio + β_6Footage + u

Table 3 displays the summaries of each of the three candidate models. Highlighted are issues present in the model, i.e., changes in significance levels. Using this table, we identify Model 3 as being a verified model and the appropriate model to answer the study. This decision was based on the model having no discrepancies in the significance levels between the training and test data set and having the highest adjusted R^2 value, compared to the other two models in the test set. Hence, the significant variables that impact the price of a condo unit in Downtown Toronto include the availability of parking, the floor level, square footage, whether the unit is a studio, the presence of a pool facility, and the age of the building. Specifically, the availability of parking, higher floor levels, and larger square footage tend to increase the price, while the presence of a studio, pool facility, and older building age tend to lower the price of a condo unit in Downtown Toronto.

Table 3 -

Summary of characteristics of three candidate models in the training and test datasets. Model 1 uses Maintenance cost, Parking, the Floor where the unit is on, Age of the condo, whether it has a Pool or Gym facility, a studio unit, Den, and Square Footage as predictors. Whereas Model 2 removes Maintenance and Den predictors from Model 1, and Model 3 removes Gym from the predictors which Model 2 has. Response is log(Price) in all three models. Coefficients are presented as estimate ± Robust SE (* = significant t-test at $\alpha = 0.05$). Issues are highlighted. Preferred model is outlined.

Characteristic	Model 1 (Train)	Model 1 (Test)	Model 2 (Train)	Model 2 (Test)	Model 3 (Train)	Model 3 (Test)
Largest VIF	4.0626266	3.207192	1.713048	1.9911653	1.7011302	1.9863066
Violations	none	none	none	none	none	none
R^2	0.8435	0.8192	0.841	0.817	0.8383	0.817
Adjusted R^2	0.8339	0.8082	0.8336	0.8084	0.8318	0.8097
Intercept	13.014 ± (0.0685) *	13.1258 ± (0.074) *	12.9757 ± (0.0642) *	13.1124 ± (0.0782) *	13.0882 ± (0.0415) *	13.1201 ± (0.0531) *
Maintenance	10^{-4} ± (10^{-4})	10^{-4} ± (10^{-4})	-	-	-	-
Parking	0.1551 ± (0.0196) *	0.1481 ± (0.0246) *	0.1539 ± (0.0198) *	0.1564 ± (0.0221) *	0.1463 ± (0.0197) *	0.1562 ± (0.0222) *
Floor	0.0011 ± (6×10^{-4})	0.0014 ± (7×10^{-4}) *	0.0013 ± (6×10^{-4}) *	0.0015 ± (7×10^{-4}) *	0.0014 ± (6×10^{-4}) *	0.0015 ± (7×10^{-4}) *
Age	-0.0153 ± (0.0016) *	-0.015 ± (0.0015) *	-0.0148 ± (0.0017) *	-0.0142 ± (0.0015) *	-0.015 ± (0.0016) *	-0.0142 ± (0.0015) *
Pool	-0.0774 ± (0.0247) *	-0.1074 ± (0.0238) *	-0.0713 ± (0.0222) *	-0.1026 ± (0.0231) *	-0.0641 ± (0.0216) *	-0.1022 ± (0.0229) *
Gym	0.1153 ± (0.0522) *	0.0032 ± (0.0527)	0.1188 ± (0.056) *	0.0079 ± (0.0573)	-	-
Studio	-0.1325 ± (0.0335) *	-0.2083 ± (0.0291) *	-0.118 ± (0.0306) *	-0.2049 ± (0.027) *	-0.1187 ± (0.0305) *	-0.2047 ± (0.027) *
Den	-0.0184 ± (0.0192)	0.0038 ± (0.0204)	-	-	-	-
Footage	7×10^{-4} ± (10^{-4}) *	7×10^{-4} ± (10^{-4}) *	8×10^{-4} ± (0) *	7×10^{-4} ± (10^{-4}) *	8×10^{-4} ± (0) *	7×10^{-4} ± (10^{-4}) *

Discussion

Based on our findings, Model 3 has been identified as the appropriate linear model to address the research question. As evident from Table 3, parking, floor, age, pool, studio, and footage are all significant predictors in this model. This conclusion can be generalized to all condo units in Downtown Toronto, beyond the sample of 315 units used in our study, as the model has been validated.

Among the significant predictors, parking, floor, and footage have positive coefficients (0.1562, 0.0015, 0.0007, respectively), indicating that an increase in floor or footage would lead to an increase in the unit's price, holding other characteristics constant. Similarly, the presence of parking would also increase the unit's price. On the other hand, being a studio unit, having a pool facility, or aging of the building are associated with negative coefficients (-0.2047, -0.1022, -0.0142, respectively), suggesting that these factors would decrease the value of the condo unit. For instance, having a pool facility on-site would lower the price of a condo unit by 10.22% (= -0.1022 x 100), assuming all other characteristics are held constant. It is important to note that since we transformed the price variable using the natural log, this is a log-level model, and the coefficient values are interpreted in percentage terms.

However, there are limitations to our study. Due to the small dataset of condo unit characteristics used, the results may not be fully representative of all condo units in Downtown Toronto, as we did not account for other factors such as on-site laundry, terrace, balcony, location of the building. If we had access to a larger dataset, like the study conducted by Kim et al., we could have conducted a comparative analysis comparing downtown condo units in different locations to identify similarities or differences in the determinants of condo unit prices in a downtown setting. Interestingly, one similarity that exists between our study and Kim et al.'s study is that the presence of a pool facility tends to decrease the price of a condo unit.

In summary, our findings suggest that parking, floor, age, pool, studio, and footage are all significant predictors in determining condo unit prices in Downtown Toronto with parking being the most influential predictor in increasing the price of a condo unit. However, it is important to acknowledge the limitations of our study due to the small dataset and the potential influence of other unaccounted condo unit characteristics. Further research with larger datasets could provide even more insights into the determinants of condo unit prices in downtown areas.

References

[1] Marcus Allen (1997) Measuring the Effects of “Adults Only” Age Restrictions on Condominium Prices, Journal of Real Estate Research, 14:3, 339-346, DOI: 10.1080/10835547.1997.12090907

[2] Kim, HG., Hung, KC. & Park, S.Y. Determinants of Housing Prices in Hong Kong: A Box-Cox Quantile Regression Approach. *J Real Estate Finan Econ* 50, 270–287 (2015). https://doi.org/10.1007/s11146-014-9456-1

Appendix

Appendix 1 –

Table showing multicollinearity between variables after transformations have been applied

	VIF	VIF > 5
Maintenance	6.681396	Y
Parking	1.906273	
Bathrooms	64.461582	Y
Floor	1.397889	
Age	2.396131	
Pool	1.523733	
Gym	1.198457	
Studio	14.348209	Y
Den	1.298920	
Bedrooms	165.287389	Y
Footage	41.790337	Y
Bathrooms: Footage	139.125320	Y
Bedrooms: Footage	251.034362	Y