What makes Toronto's expensive neighborhoods expensive?

Beiqin Zeng, Sarah Hafez, Wen Li

December 15th, 2021

Introduction

Over the last year, Toronto's housing prices rose by over 40%. In fact, over the last few years, housing prices have faced a steady increase by 10% each year (Gurney, 2021). To many, there aren't many tangible justifications behind these prices. In this project, we look at some of Toronto's housing data to see if it can offer any explanations. Specifically, we focus our investigation towards Toronto's twenty-two neighborhoods, with the goal of answering the question: *"What makes Toronto's most expensive neighborhoods expensive?"*. To answer this question we have scraped data about Toronto's sold condos from October 2020 to October 2021 from condos.ca. Using this data, we have built a multi-variable linear regression model to draw comparisons between Toronto's expensive and less expensive neighborhoods. We discover that on average, a 800 sqm condo costs an extra $85,600 in the expensive neighborhoods, an amount that represents ~10% of the average condo price. The data points to the looser pet restrictions, newer buildings, and the lower number of children as some of the advantages in the expensive neighborhoods. However, one must be careful when treading the line of correlation vs causation.

The Data

A. Data Collection: Methods and Overview

To collect our data, we scraped condos.ca, which has the largest database of all condo units in Toronto and offers users detailed market analysis specific to each building in Toronto. We collected the data for all the sold condos in downtown Toronto from October 2020 to October 2021.

We used *Request* and *BeautifulSoup* libraries in Python and *jsonlite* library in R as well. The main scraping tasks were divided into three subtasks:

1. Obtaining the full list of condos in downtown Toronto that have been sold in the time period specified;
2. Obtaining the detailed information for each listing, for example, the condo size, the location, the number of bedrooms, etc.;

3. Obtaining the demographics information for the twenty-two neighborhoods in downtown Toronto. For each area, we collected details such as: the number of schools, the age distribution of the population, etc.

After scraping the data, we ended up with more than 8600 original data points. After removing the duplicates and non-condo properties, about 2100 condos are left. Table 1 shows the data collected for all condos.

Table 1: Variables found in the condos data

address	price	number of bedrooms	number of bathrooms
number of parking spots	property tax 2021	actual size	exposure
maintenance fees	age of building	outdoor space	locker
heating type	parking type	property type	neighborhood
ensuite laundry	corporate number	size range	amenities

Table 2 shows the data collected for all neighborhoods.

Table 2: Variables found in the demographics data

average income	tenancy and property type	children per household	commute type
household composition	languages	educational levels	age distribution

B. Data Exploration: Toronto's Most Expensive Neighborhoods

The first task after obtaining the data was to define "expensive areas". We defined it as the top five neighborhoods with the highest *price per squared meter(sqm)* values instead of *price*. We chose to work with the *price per sqm* as some condos are expensive by virtue of their size, which is not an interesting finding and we are interested in a more complex explanation of the higher prices. Figure 1 also demonstrates the difference in distributions between the *price* (left) and *price per sqm* (right). The distribution of *price* has very long right tail, indicating the presence of many and sizable outliers. Many condos have a price of between $500,000 to $1,500,000, while some have exceed the 4 million dollar mark. When we divide by the condo size, the *price per sqm* distribution resembles a normal distribution with a slightly longer right tail.

We also found that rankings when it came to pure prices were deceiving. For example, Annex-UofT was the third most expensive area when it came to price. However, this was came with a caveat: it had the third most spacious condos in Toronto. As such, when we divided by the condo size, Annex-UofT was shown to be one of Toronto's least expensive neighborhoods, with it ranking as third to last when it comes to *price per sqm*. Appendix A shows the precise average price, price per sqm and condo size in all neighborhoods.

Finally, according to the *price per sqm* rankings, Toronto's most expensive neighborhoods came to: Yorkville, Queen West, Grange Park, Yonge and Bloor, and The Waterfront.

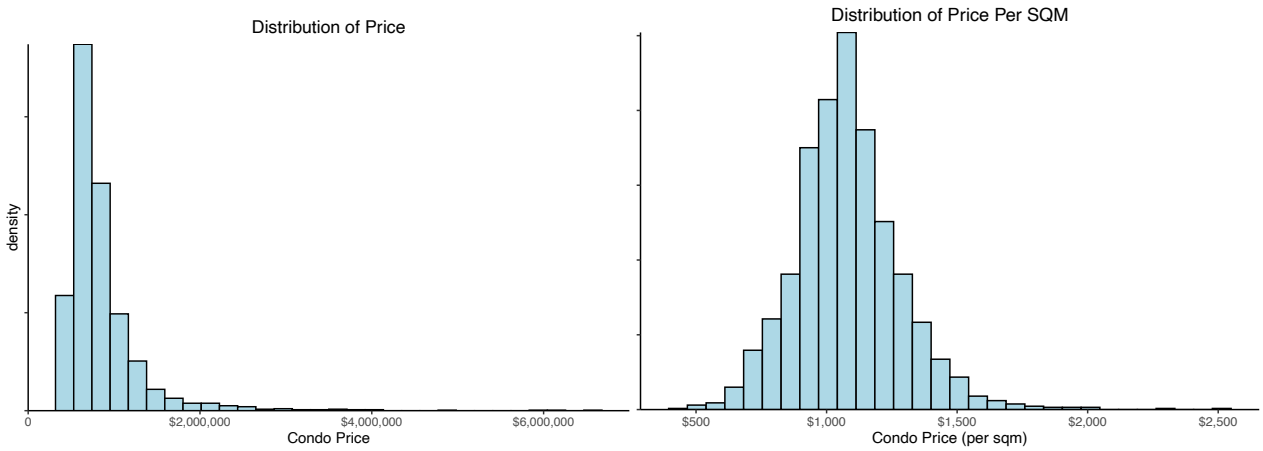

Figure 1: Distribution of Price and Price Per SQM

Table 3: Differences in Price between the Different Neighborhoods

Avg. Price Per Sqm (expensive)	**Avg. Price Per Sqm (other)**	**Avg. Condo Size**	**Avg. Condo Price**	**Price Difference (adjusted to condo size)**
$1,159	$1,052	800 sqm	$853,182	$85,600

Table 3 shows that Toronto's most expensive neighborhoods have a $107 *upcharge* per sqm in comparison to other neighborhoods. This means that an average condo with a size of 800 sqm is $85,600 more expensive in the aforementioned neighborhoods. The average condo price is around $850,000. This $85,600 upcharge represents 10% of the average condo price, and are as such by no means negligible. We also ran a Wilcoxon test to find whether the differences in *price per sqm* were statistically significant, which was indeed the case. With this information at hand, it becomes more intriguing to find what makes the expensive neighborhoods have such a significant upcharge.

Figure 2 shows the *price per sqm* distributions in each neighborhood. As well, the neighborhoods are ordered from the least to the most expensive, making our expensive areas of choice the last five areas on the right. Interestingly, it seems like there are small differences between the median *price per sqm,* with all neighborhoods ranging from $900 to $1,100. This is true but one should keep in mind that these difference become very large when we consider the actual size of the condo. A $100 difference in median means a $80,000 difference in price on average.

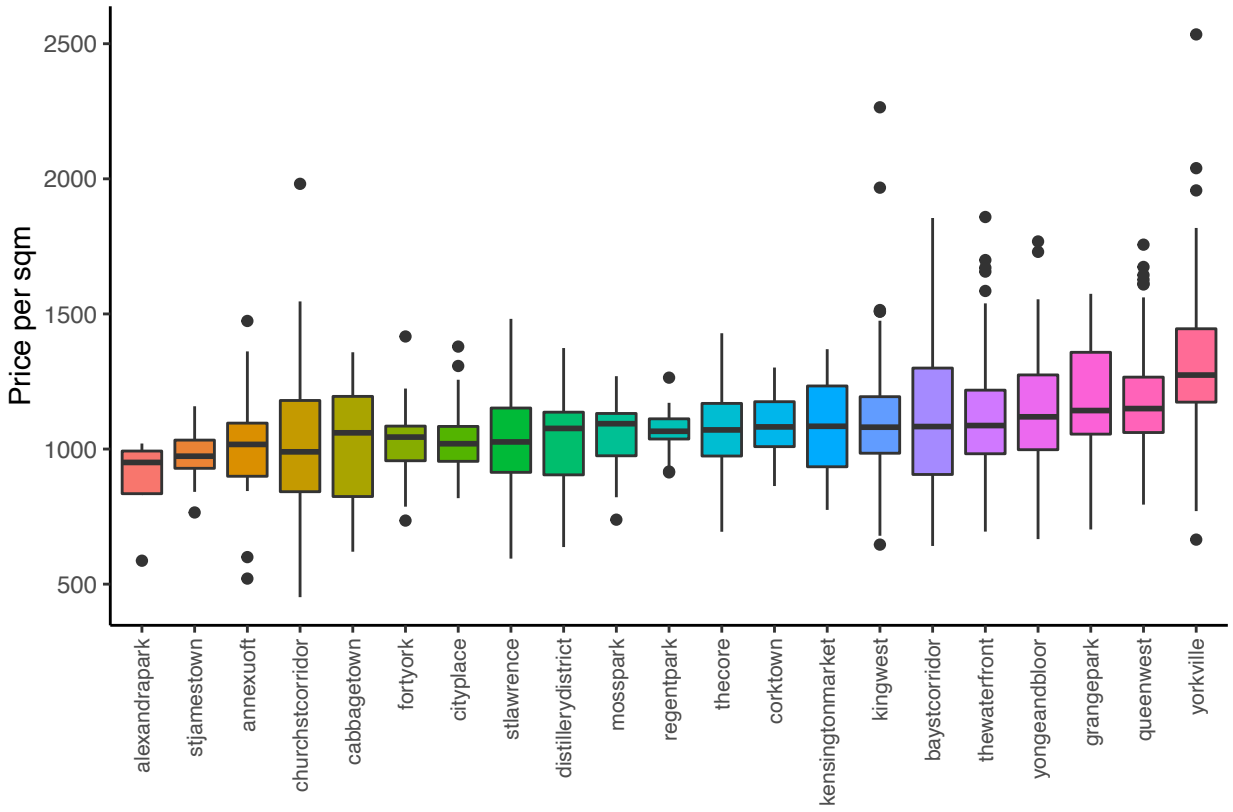

Figure 2: Price Distribution by Neighborhood

Analysis: Methodology

Our methodology can be broken down to two parts:

1. Using the linear regression model to determine the relationship between all factors and condos' price;
2. Determining which factors have significant differences between the expensive neighborhoods and the less expensive neighborhoods, using Wilcoxon Test. As well as, determining the effect of these differences.

Feature Selection and Data Preprocessing

Before building the linear regression model, some considerations had to be made for the data:

1. Some factors are dependent on the price. For example, the condos data included information on the property tax which are a direct result of the condo price. As such, we removed all tax data from our model. The demographics data included information on the average individual and household income levels in each neighborhood. People with high levels of income likely live in expensive neighborhoods. In a way, the income level is an easy predictor of price. In order to achieve a more complex model, we removed all indicators of income from the set of explanatory variables.

2. For each area, condos.ca provided a breakdown of the neighborhood's statistics. In our dataset, the breakdown of each statistic has its own statistic. For example, the household composition is expressed

through four variables: Single Family, Multi Family, Singe Person and Multi Person. Naturally, these variables add up to 1. To avoid having multicollinearity in our data, it was important to drop at least one of these variables for each neighborhood statistic.

3. Additionally, the neighborhood statistics were represented in decimal place. For example, the 40% single family household composition in Alexandra Park was represented as 0.4. We multiplied these variables by 100 to reach an easier interpretation of the model. This way, it will be easier for the untrained eye to see that if the composition of single family increases by 1% (one unit of change), the price will increase by the coefficient amount.

4. There are many variables in the data. Not only so, but many variables were categorical and included many levels, bringing up the total to 250 variables. For example, we had information on the amenities offered in each condo building, with a total of 76 amenities. We wanted our model to be as simple as possible, as such, we narrowed down the 76 amenities into 13, such as availability of in-building gyms, patios, meeting rooms and concierge. Following the same logic, we removed columns that we didn't perceive as interesting to the price, such as: maintenance fees (a $1000 maintenance fee doesn't deter someone from buying a half million dollar condo), and the breakdown of housing types in each neighborhood.

5. We also used Akaike information criterion (AIC) with backwards selection to select the best model. This helps improve the fit of the model as well as its specificity.

Results

The Linear Regression Model: After the initial preprocessing steps, we ended up with a model that achieves a R^2 score of ~0.5, which indicates the model is only able to explain 50% of the variability in prices. This isn't necessarily bad given the limitations of the data we have at hand as well as the limitations of the linear assumption. Appendix B containes the full regression results for our model. Over the following section, we go deeper into the linear regression model results.

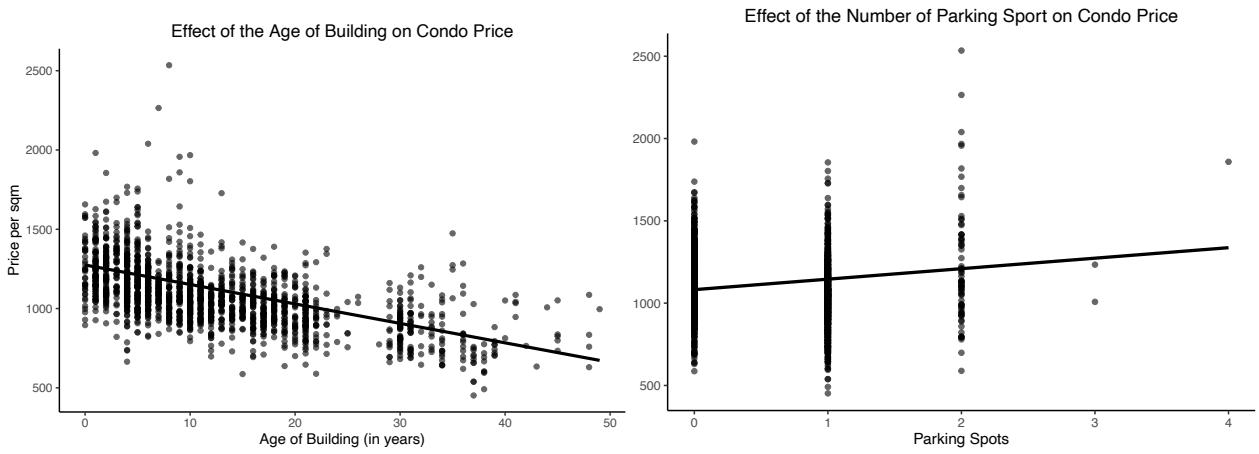

Figure 3: Effect of Building Age and Parking Spots on the Price per sqm

Interpretations of the Linear Regression model: The model indicated that the age of the building has a negative effect on the *price per sqm.* When we look at figure 3 (left plot), we can see that the data supports that claim. Older buildings (30 years+) seem to have a smaller and more constrained range of prices in comparison to the newer buildings. The model also indicated that the number of parking spots has a positive effect on the *price per sqm.* The right plot doesn't seem to support that claim. Condos with no parking spot had a *price per sqm* ranging from $600 to $1,600. Similarly, condos with two parking spots also stuck to the same range, barring some outlier condos. After looking more deeply into the data, it is hard to believe that an extra parking spot caused for a $64 in *price per sqm* as the model indicated. This also supports the fact that the model doesn't seem to fit well to the data.

We aren't only interested in factors with statistically significant p-values (which is the case for almost all variables), but more specifically, we are looking for variables with huge effects and large coefficients. For example, when looking at fr_only, which describes the percentage of french-only speakers, the model suggests that an increase in that percentage by one increases the price per sqm by more than 34,000. However, this number is misleading as all neighborhoods have a really small proportions for french only speakers, sitting at a low 0.09%. This is also the case for mutli_family, which describes the percentage of condos with more than one family living in that space. The coefficient is more than $30,000 which is also met with the caveat that only 0.07% of household live under such condition.

The model also suggests that condos with an East-West exposure cost an extra $200 per square meter in comparison to those with other exposures. This is a sizable cost as the average price per sqm is ~$1,000. However, we found that only 7 observations have such exposure, which makes it difficult for us to take that result at face value. The model was seemingly highly influenced by these observations.

The linear regression model included upwards of 30 variables. Some of the interesting insights that we have found is that pet restrictions and an older building have a negative effect on the price while having a security guard and a gym in the building has a positive effect on the price. Over the next section we go over the differences between the expensive and the less expensive neighborhoods.

What makes Toronto's expensive neighborhoods expensive? Seeing the significant factors in the linear regression model leads us to the following question: are the expensive and non-expensive neighborhoods any different when considering these factors? To be more specific, are these differences statistically significant when considering the Wilcoxon test? Table 4 contains the answers to these questions.

Table 4: Qualitative Differences between Neighborhoods

Factor	**Are the expensive neighborhoods statistically different than the non expensive neighborhoods?**
Pet Restrictions	Yes. Expensive neighborhoods have less pet restrictions.
Security Level	No. All neighborhoods have good security measures.
Gym Availability	No. All neighborhoods contain a fair amount of gyms in the building.
Age of Building	Yes. Expensive areas tend to have newer buildings.
Number of Children Per Household	Yes. Expensive neighborhoods have on average a lower number of children per household.

Table 5: Qualitative Differences between Neighborhoods

Variable	Expensive Neighborhoods	Less Expensive Neighborhoods	Unit Effect on Price/Sqm	Expensive Neighborhood Upcharge
Avg. Number of Children	0.51	0.6	-5,142	$430
Avg. Age of Building	11.3	12.7	-12	$17
Buildings with Pet Restrictions	54.6%	65.4%	-67	$723

Table 5 quantifies some interesting differences between the expensive areas and less expensive areas. It also quantifies the effect of these differences on the price per square meter. The first column is the average value in the expensive neighborhoods, and the second column is the average value in the less expensive neighborhoods. The third column is the variable's coefficient value from the fitted model which is defined as the unit effect on the price per sqm. The fourth column is the result of multiplying the third column by the difference between the first two columns. It represents the extra money that expensive areas charge you for these differences.

When we interpret the results in table 5, we should be careful of the fact that the average *price per sqm* is only ~$1,000. It would then be unreasonable to suggest that the average number of children caused for a $430 upcharge in the expensive neighborhoods or that the looser pet restrictions caused for a $723 upcharge; both numbers combined exceed $1,000 and exceed the original $107 upcharge that we have specified. It is likely the case that the $107 upcharge is a result of a mix between an *upcharge* for some positive factors (some we know and some we don't) as well as a downcharge for some negative factors. What the data tells us is that the average number of children, pet restrictions and age of buildings seem to play a positive role towards the aforementioned upcharge.

Conclusion

The goal behind this project was to find if there are any explanation behind some of Toronto's pricey neighborhoods. Specifically, we aimed to answer the question: "What makes Toronto's expensive neighborhoods expensive?". We used a linear regression model that pointed our focus towards some variables and explored whether these variables have significant differences between the expensive and less expensive neighborhoods. We found that expensive neighborhoods enjoy a lower number of children, looser pet restrictions and newer buildings. As well, these factors are highly correlated with the higher price tag. However, the definite answer to the question likely boils down to more complex factors.

A dilemma that we have met is whether our variables are a reason or a result of the higher prices. One could argue that a lower number of children can lead to higher prices as it makes the neighborhood less "busy" and "loud". However, research also shows that higher-income families tend to have a lesser number of children. One could then argue that a lower number of children is a result of the neighborhood being expensive. It isn't as clear cut as some might think.

The data we obtained had its limitation. In hindsight, it would have been useful to have access to data about the distance to the city center or the subway, information about the air quality or even data about the construction in the neighborhood. We also acknowledge that there can be unquantifiable variables affecting the price, such as popularity; there is a possibility that some neighborhoods suddenly rise in popularity which affects the real estate market.

To conclude, it is clearly the case that the linear regression model wasn't able to capture the complex structure of the housing market. Nonetheless, we were able to obtain some interesting insights using it. A good extension to this research is to explore a hierarchical model such as decision trees to capture the more complex structure of the neighborhoods.

References

Gurney, M. (2021, August 16). *The quick fix, Part 1: How Ontario can improve its housing situation — now*. Retrieved from TVO: https://www.tvo.org/article/the-quick-fix-part-1-how-ontario-can-improve-its-housing-situation-now

Appendix

Appendix A: Average Price, Price per SQM and Condo Size in all Neighborhood

location_area	mean_size	mean_price	mean_price_per_sqm
alexandrapark	810	674625	877
annexuoft	961	994204	1007
baystcorridor	822	878818	1107
cabbagetown	680	643317	1020
churchstcorridor	726	681482	1010
cityplace	755	763568	1025
corktown	650	698869	1085
distillerydistrict	772	784156	1036
fortyork	825	830233	1023
grangepark	758	856050	1171
kensingtonmarket	686	716438	1087
kingwest	860	934317	1090
mosspark	665	684028	1062
queenwest	756	879188	1173
regentpark	658	691322	1067
stjamestown	797	771467	978
stlawrence	814	818929	1028
thecore	715	764608	1079
thewaterfront	974	1067790	1111
yongeandbloor	717	834530	1163
yorkville	1046	1459518	1327

Appendix B: Final Linear Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8665	2670	3.245	0.001197
num_of_bed1+1	-49.41	10.27	-4.811	1.629e-06
num_of_bed1+2	-153.8	83.34	-1.845	0.06517
num_of_bed2	-72.52	10.95	-6.62	4.762e-11
num_of_bed2+1	-86.26	12.55	-6.874	8.662e-12
num_of_bed2+2	-107.4	83.61	-1.284	0.1993
num_of_bed3	-92.77	17.5	-5.301	1.296e-07
num_of_bed3+1	-206.8	37.07	-5.578	2.809e-08
num_of_bed3+2	27.59	143.9	0.1917	0.848
num_of_bed4	127.9	106.3	1.203	0.229
num_of_bed4+1	-153.6	143.5	-1.07	0.2847
num_of_bedStudio	52.03	12.71	4.094	4.428e-05
num_of_parking	64.03	7.83	8.177	5.498e-16
age_of_building	-12.36	0.4007	-30.86	1.522e-167
include_security_guard	35.81	9.093	3.939	8.515e-05
include_pet_restriction	-66.99	8.203	-8.166	6e-16
include_gym	41.38	16.08	2.574	0.01013
exposure_EW	220.5	49.98	4.412	1.085e-05
transit	-2358	394.2	-5.983	2.642e-09
foot	-2472	412.5	-5.992	2.508e-09
bicycle	-2411	399.4	-6.037	1.915e-09
drive	-3361	562.3	-5.977	2.752e-09
single_family	-346.4	59.19	-5.853	5.733e-09
multi_family	30330	5116	5.928	3.672e-09
single_person	449.4	76.12	5.903	4.267e-09
owners	779.7	130.5	5.973	2.818e-09
en_only	5086	852	5.969	2.874e-09
fr_only	34547	5820	5.936	3.519e-09
en_and_fr	1793	304.6	5.886	4.732e-09
no_high_school	-2501	428.7	-5.834	6.418e-09
high_school	-3360	569.8	-5.897	4.416e-09
college_certificate	-2164	366.4	-5.906	4.202e-09
university_certificate	-995	178.1	-5.587	2.667e-08
bachelor_degree	-3445	582.9	-5.91	4.106e-09
post_graduate_degree	-1756	302.6	-5.803	7.719e-09
avg_children_per_household	-5141	830.4	-6.191	7.442e-10
apartment_1_to_4_floors	205.6	32.89	6.25	5.133e-10

Table 8: Fitting linear model: price_per_sqm ~ num_of_bed + num_of_parking + age_of_building + include_security_guard + include_pet_restriction + include_gym + exposure_EW + transit + foot + bicycle + drive + single_family + multi_family + single_person + owners + en_only + fr_only + en_and_fr + no_high_school + high_school + college_certificate + university_certificate + bachelor_degree + post_graduate_degree + avg_children_per_household + apartment_1_to_4_floors

Observations	Residual Std. Error	R^2	Adjusted R^2
1799	142.9	0.5117	0.5017