Building Integrity in Retrospect: Assessing the Impact of Age on 2023 Safety Evaluations*

Terry Tu, Jingyi Shen, Yaning Jin

March 11, 2024

This study harnesses a dataset from the Toronto Open Data portal, comprising 3,435 records of apartment evaluations, to analyze the correlation between building age and safety scores. Employing a linear regression model, we discovered a significant inverse relationship, highlighting that older buildings tend to have lower safety evaluations. The findings suggest a need for policy reform, particularly in the maintenance and inspection of Toronto's aging residential infrastructure. Despite the study's focus on building age, further research considering additional variables could offer a more comprehensive understanding of factors influencing building safety.

Table of contents

---

*Code and data supporting this analysis are available at https://github.com/TEJMaster/Toronto-Building-Safety-Analysis-2023.git

1 Introduction

In the wake of the catastrophic collapse of an aged condominium in South Florida in June 2021, which resulted in 98 casualties, a subsequent lawsuit by the bereaved families underscored the notion that routine inspections and maintenance could have forestalled the disaster(Anderson 2021). This incident precipitates our investigation into forecasting the building evaluation scores for Toronto residences in 2023 via a simple linear regression model. An analysis was conducted on the apartment building evaluation dataset obtained from the City of Toronto's open data portal to assess the relationship between the age of a property and its evaluation score assigned by the city. Our research is singularly focused on elucidating the effect of one independent variable—building age—on the dependent variable, which in this context is the building evaluation score.

Using simple linear regression analysis enables a detailed exploration of the linear relationship between the age of a building and its evaluation score. By quantifying this relationship through regression coefficients, we aim to reveal the expected change in evaluation scores for each incremental year in building age. The statistical significance of the age variable on evaluation scores will be ascertained using p-values. Provided the model satisfies the underlying assumptions, t-tests will be applied to test the hypotheses, while confidence and prediction intervals will be leveraged to quantify the effect size and enhance the accuracy of our forecasts, respectively.

2 Data

In this research, we examine apartment data sourced from the Toronto Open Data portal. The dataset encompasses 3,435 records, focusing on specific variables: the ward of each apartment, its evaluation score, and the construction year. The data covers evaluations across Toronto's 25 wards, with scores varying between a minimum of 0 and some achieving the maximum score of 100. The study includes apartments that are over three stories in height and contain at least 10 units. These apartments span a wide range of construction years, from as early as 1805 to as recent as 2023. All the apartments in this dataset adhere to the City's Apartment Building Standards bylaw, which is designed to safeguard the interests and safety of both landlords and tenants by mandating these evaluations (Toronto 2023).

2.1 Variable Description

CURRENT BUILDING EVAL SCORE: The Current Building Evaluation Score represents a measure of a building's adherence to property standards. This score is derived by summing two components: the current reactive score, which reflects the compliance based on any outstanding Orders and Notices of Violations, and the proactive building score, which is based on the most recent comprehensive evaluation of the building.

YEAR BUILT: This indicates the year when the building was initially constructed, sourced from Building Owners/Managers.

YEAR EVALUATED: This denotes the year in which the building underwent evaluation, reflecting its condition and performance.

AGE: In this research, neither YEAR BUILT nor YEAR EVALUATED was used directly, we have generated a new variable called AGE which use YEAR EVALUATED minus YEAR BUILT.

2.2 Data Analysis Tools

The data analysis was performed using R (R Core Team 2022), a powerful open-source statistical programming language. Key packages from the tidyverse collection (Wickham et al. 2019) were employed to streamline data manipulation, visualization, and analysis processes. These packages include ggplot2 (Wickham 2016) for creating advanced graphics, dplyr (Wickham et al. 2022) for data manipulation, readr (Wickham, Hester, and Bryan 2022) for its robust data reading functionality, here (Müller 2020) is used to avoid file path issue, and knitr (Xie 2014) for dynamic report generation.

2.3 Data Exportation

After cleaning, the modified dataset, now stored in the `cleaned_data` data frame, is written to a new CSV file named 'cleaned_building_data.csv'. This file is saved in the 'outputs/data' directory, which is again constructed using the `here` function for path management. Saving the cleaned dataset allows for a stable version of the dataset to be used for analysis or sharing, ensuring reproducibility of results.

Statistically, this script is preparing the dataset for analysis by ensuring the integrity and relevance of the data, which is foundational for any reliable statistical analysis. By creating the 'AGE' column, the script enables researchers to analyze the relationship between the age of a building and other variables of interest, such as compliance with property standards, in a quantitative manner.

3 Model

4 Result

4.1 Coefficient

```
      Estimate Std. Error    t value     Pr(>|t|)
```

```
(Intercept) 93.6218207 0.64813240 144.448605 0.000000e+00
AGE         -0.1008607 0.01015308  -9.934001 1.166964e-22
```

Intercept: The intercept β_0 an estimated value of 93.62182. This is the expected value of the response when the independent variable x is zero. The standard error of the intercept is 0.64813, indicating the average distance that the observed values fall from the regression line.

AGE: The coefficient for AGE is -0.10086 which in terms of β_1 is -0.10086. This suggests that for each additional year of age, the response decreases by approximately 0.10086 points, on average. The standard error of this estimate is 0.01015, which is a measure of the precision of the estimate.

T-value: The t-value is a measure of how many standard deviations the coefficient is away from zero. For the intercept, the t-value is 144.449, and for AGE, it is -9.934. These are calculated by dividing the estimate by the standard error. The high absolute value of the t-statistics suggests that the coefficients are significantly different from zero.

P-value: The p-value tests the null hypothesis that the coefficient is equal to zero. A p-value less than 0.05 typically suggests that there is a statistically significant association between the predictor and the response variable. In this table, both p-values are shown as "2e-16", which is a notation for a number very close to zero and indicates that both the intercept and the AGE predictor have p-values much lower than 0.05, making them highly significant.

In summary, the linear regression analysis suggests that there is a statistically significant inverse relationship between the age of an apartment and its evaluation score. The older the apartment, the lower its evaluation score tends to be, holding other factors constant. The intercept indicates that a newly built apartment (AGE = 0) would have a predicted evaluation score of 93.62182.

4.2 Residuals

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-69.662  -4.364   1.702   0.000   5.833  15.859
```

1. Minimum Residual: This value suggests the presence of at least one observation where the building evaluation score was underestimated by the model by as much as 69.662 points, given the age of the building. This could be an outlier or may indicate a building whose score is exceptionally low compared to what the model predicts for its age.

2. First Quartile Residual: 25% of the residuals are less than this value. It implies that a quarter of the buildings have evaluation scores that are 4.364 points less than the predicted score by the model for buildings of their respective ages.

3. Median Residual: The median of the residuals is slightly above zero, indicating that the distribution of errors is not perfectly symmetrical and that there may be a slight tendency for the model to overestimate the scores, on average, for the buildings in this sample.

4. Third Quartile Residual: This is the median of the upper half of the residuals. 75% of the buildings have evaluation scores within 5.833 points of the model's prediction or better, reflecting that for the majority, the model's predictions are quite close to the actual scores.

5. Maximum Residual: The largest positive residual suggests that the model overestimated the building evaluation score for at least one building by as much as 15.859 points compared to its actual score. This again may indicate an outlier or a building that performed unexpectedly well despite its age.

These residuals are crucial for assessing the adequacy of the model fit and the validity of its assumptions, such as linearity, homoscedasticity (constant variance of residuals), and normality of error terms. The range and distribution of these residuals, particularly the presence of significant outliers, can influence the conclusions drawn from t-tests and the calculated confidence and prediction intervals. Such anomalies warrant further investigation into the specific characteristics of those buildings to understand why the model's predictions differ markedly from the observed scores. It's also important to ensure that the regression assumptions hold since this affects the validity of the p-values and confidence intervals used to quantify and infer the effect of building age on evaluation scores.

4.3 Plot for the linear model

Relationship between Building Age and Evaluation S
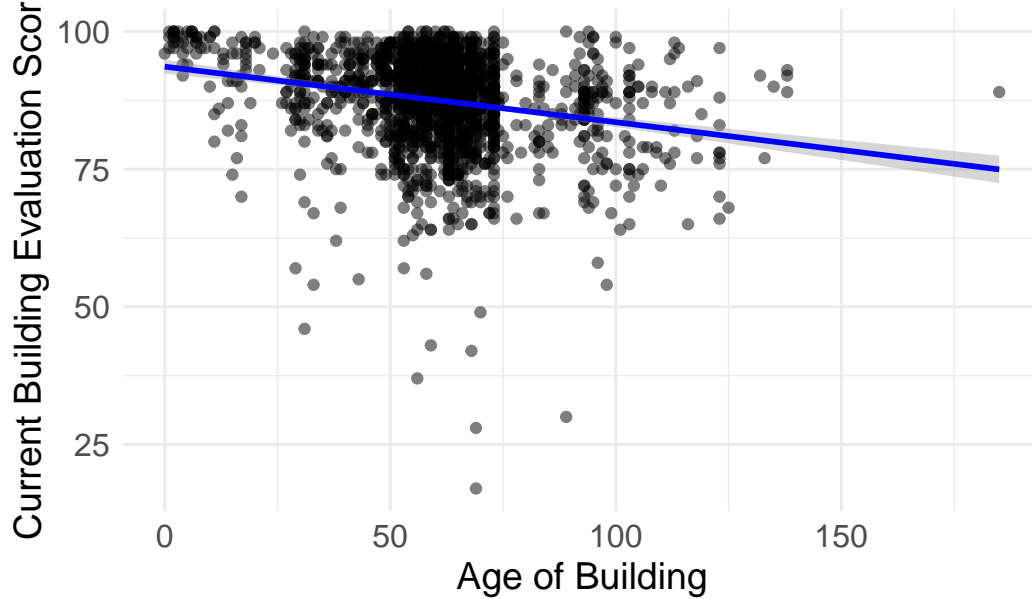
The graph provided illustrates the findings of a simple linear regression analysis that seeks to clarify the influence of one independent variable, building age, on the dependent variable, the building evaluation score. The scatterplot displays individual data points representing apartments, with the age of the building on the x-axis and the current building evaluation score on the y-axis.

A linear regression line has been fitted through the data, represented by the solid blue line, which indicates the average trend in the data. The slope of this line is a graphical representation of the regression coefficient associated with building age. In statistical terms, the slope tells us the average change in the evaluation score for each year increase in building age. If the slope is negative, it would suggest that, on average, older buildings tend to have lower evaluation scores. The density of the data points is heaviest near the younger age of buildings, and as age increases, the data points disperse more. This could suggest that there is greater variability in the scores of older buildings or that there are simply fewer older buildings in the dataset.

Statistical significance is often determined by looking at the p-value for the slope of the regression line. In previous discussions, a p-value well below the threshold of 0.05 was mentioned, indicating that the effect of building age on evaluation score is indeed statistically significant. This would mean that there is a very low probability that the observed relationship, or one more extreme, would occur if there were actually no relationship between building age and evaluation score in the population.

It is also important to examine the residuals, or the differences between the observed scores and those predicted by the model, which should be randomly dispersed without pattern if the model's assumptions hold true. In addition, the graph should be assessed for homoscedasticity, meaning that the variance of the residuals is constant across all levels of the independent variable. If the residuals fan out or form a funnel shape as building age increases, this would indicate heteroscedasticity, a violation of one of the linear regression assumptions, which could affect the reliability of the model.

To summarize, this graph presents a visual and statistical summary of the relationship between building age and evaluation score, with the analysis revealing that age is a significant predictor of evaluation score, though it accounts for a relatively small portion of the variability in scores. Further analysis would be necessary to better understand the dispersion of scores, especially among older buildings, and to examine the model's underlying assumptions.

5 Discussion

5.1 Overview of the Study

This paper presented a detailed analysis using simple linear regression to investigate the impact of building age on the evaluation scores assigned by the City of Toronto. By analyzing data from the Toronto Open Data portal, the study aimed to forecast evaluation scores for Toronto residences in 2023. The analysis revealed a statistically significant inverse relationship between building age and evaluation score, suggesting that older buildings tend to have lower evaluation scores.

5.2 Insights into Urban Development and Maintenance

One key insight from this study is the importance of building age in the overall health and safety of urban residential environments. The findings underscore the potential risks associated with older buildings, which may not meet modern standards or may have deteriorated over time. This knowledge emphasizes the need for regular inspections, maintenance, and upgrades to older buildings to ensure they remain safe and habitable. Additionally, the study highlights the value of continuous monitoring and evaluation by urban authorities to maintain and improve housing standards across the city.

5.3 Policy Implications and Housing Quality

Another significant insight relates to the policy implications for housing quality and safety standards. The correlation between building age and lower evaluation scores may prompt policymakers to reconsider regulations surrounding building inspections, maintenance schedules,

and renovation requirements. It suggests an opportunity for targeted interventions in older buildings to prevent potential hazards and improve living conditions for residents. Moreover, these findings could guide urban planning strategies, prioritizing the revitalization of aging housing stock and ensuring equitable access to safe and quality housing.

5.4 Limitations of the Study

Despite its contributions, this study has several limitations. Firstly, the model accounts for only a small fraction of the variability in evaluation scores, indicating that other factors besides building age significantly influence these scores. The exclusion of variables such as building materials, design, maintenance history, and location may limit the comprehensiveness of the findings. Additionally, the model's assumptions (e.g., linearity and homoscedasticity) and the presence of outliers suggest caution in generalizing the results without further investigation.

5.5 Future Directions

Given the limitations and the insights gained, future research should aim to incorporate additional variables that could affect building evaluation scores. A more comprehensive model including factors like renovation history, compliance with current building codes, and socio-economic characteristics of the neighborhood could provide a deeper understanding of what influences building evaluation scores. Longitudinal studies could also shed light on the trends over time, offering insights into the effectiveness of policy interventions and maintenance practices. Moreover, comparative analyses across different cities or countries could highlight unique challenges and best practices in building maintenance and safety standards.

In conclusion, while this study has provided valuable insights into the impact of building age on evaluation scores, it also opens the door for further research to explore the multifaceted nature of building quality and safety. Addressing these questions is crucial for developing effective strategies to improve urban living conditions and ensure the longevity and safety of the housing stock.

References

Anderson, Curt. 2021. "Lawsuit Claims Surfside Condo Collapse Triggered by Building Work." *NBC 6 South Florida.* NBC 6 South Florida. http://www.nbcmiami.com/news/local/lawsuit-claims-surfside-condo-collapse-triggered-by-building-w ork/2622574/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Toronto, City of. 2023. "Open Data." *City of Toronto*. https://www.toronto.ca/city-government/data-research-maps/open-data/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.