Corrrelation Between Location and Building Evaluation Score*

Hyuk Jang

2024-01-23

This paper analyzes the correlation between location of apartment buildings in Toronto and its evaluation score. This will help us analyze some safer and cleaner neighberhoods and what may be the reason why such correlation occurs.

1 Introduction

The demand for housing in Toronto is very high while supply is comparatively lower. This has led to soaring housing prices and house buyers may have lots of concerns regarding which and where to purchase houses. This paper will create an analysis of how location can affect building evaluation score and which area has the highest average score.

To analyze the correlation, this paper is divided into the following subsections: data, results, discussion, and conclusion. The data section will discuss the information gained from cleaning the raw data. The result section will further deeply analyze trends. Conclusion will summarize the paper and give further insight.

2 Data

The data used in this paper is obtained from the Open Data Toronto Portal, using the 'opendatatoronto' library (**rOpenDataToronto?**). The data set used is named 'Apartment Building Evaluation' (Data Pre 2023). The data in the section is obtained by using the program (**r?**), with support from `tidyverse` (**rTidyverse?**), `ggplot2` (**rGgplot2?**), `dplyr` (**rDplyr?**), `readr` (**rReadr?**), `tibble` (**rTibble?**), `janitor` (**rJanitor?**),`KableExtra` (**rKableExtra?**), `knitr` (**rknitr?**), and `here` (**rHere?**).

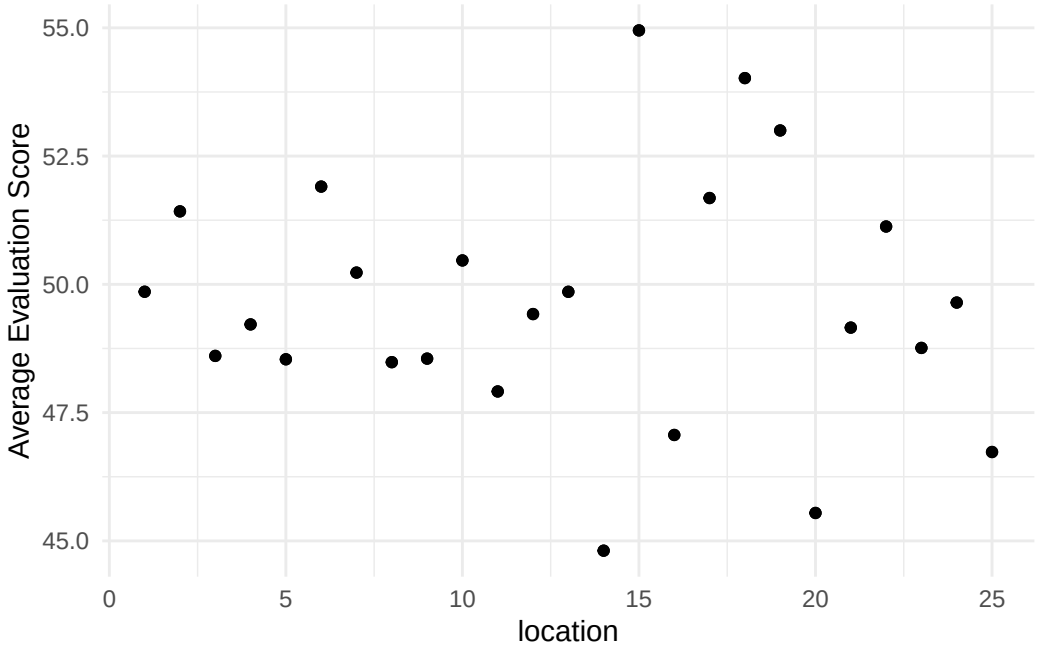
*Code and data supporting this analysis is available at: https://github.com/anggimude/Correlation-Between-Location-and-Building-Evalualtion-Score.git

We are interested in the correlation between location (ward) and the average score of each ward. This is the purpose of the paper is to analyze if there are specific locations that have better quality apartments than others do.

To accomplish the task, the scripts that are written provide us a table of the sum of all scores and the average score of each location. There is a total of twenty-five wards in Toronto and the scores vary from a minimum of 0 to a maximum of 100.

3 Results

```
clean_evaluation_data |>
  ggplot(mapping = aes(x = Ward, y = Average_Score)) + geom_point() + geom_beeswarm() + la
```



Based off of the scatter plot that is made, we can finally analyze the data. The plot shows us that there necessarily isn't a direct correlation between location and the apartment building evaluation scores. However, we can find some patterns that do appear but not trends. Some of the patterns that are observed is that from the location (ward) 1~13, the average scores tend to be around 50 or slightly lower than so. The minimum average score is about 45, and the maximum is 55. This tells us that the average score of all the apartments that have been

evaluated Pre 2023 doesn't necessarily have an outlier and mostly have a score lower than 50.

4 Discussion

There are several limitations to this analysis. Firstly, we do not know exactly how the score is rated and how strict they are on the grading which makes it difficult for us to know the standard score and compare it to an apartment in interest. Another is that I didn't take into account the factors of proactive building score nor the current reactive score. Taking into account these data and further deeply analyzing may give us a better insight. Lastly, the data that is used in this paper is Pre2023, and each building must undergo evaluation at least once every two years, which makes some of the data outdated as it is 2024 now.

5 Conclusion

This paper has investigated the correlation between location and apartment building evaluation score within Toronto Pre 2023. The data analysis has shown that there isn't a strong correlation, there does exist a weak correlation where the score is averaged at slightly lower than 50. Thus, we do not have strong evidence that certain neighborhoods have cleaner or better evaluated apartments compared to other neighborhoods as they are all quite similar. In the future, if someone was to reference the data before move-in or a new purchase it may be more accurate to look at the most recent data and instead of looking at the average score of a ward, looking at the specific apartment can be more helpful in making a decision. Further analysis would be needed with an improvement in coding skills.

References

Firke, Sam. 2021. Janitor: Simple Tools for Examining and Cleaning Dirty Data. https://CRAN.R-project.org/package=janitor.

Gelfand, Sharla. 2022. Opendatatoronto: Access the City of Toronto Open Data Portal.https://CRAN.Rproject.org/package=opendatatoronto.

Müller, Kirill, and Hadley Wickham. 2022. Tibble: Simple Data Frames. https://CRAN.R-project.org/package=tibble.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.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. 2021. 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/9781466561

Zhu, Hao. 2021. kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax. https://CRAN.R-project.org/package=kableExtra.

Zhu, Hao. 2021. kableExtra: Construct Complex Table with Kable and Pipe Syntax. https://CRAN.R-project.org/package=kableExtra.