A Statistical Analysis of Toronto Residents' Engagement and Their City Impression.*

Exploring the Correlation Between Engagement and Perception.

Nikhil Iyer

January 25, 2024

Cities go through great lengths to make sure their residents are content, and Toronto is no different. Year after year Toronto is listed as one of the best cities in the world, so how do the people actually living in the city perceive it? Through statistical analysis of survey's conducted by Toronto, the top-level finding was that the overall sentiment did not change, in fact it slightly decreased as resident engagement increased.

1 Introduction

There has been an exponential growth in the number of people living in urban areas than rural, and for the first time, more people live in urban than rural areas (United Nations Department of Economic and Social Affairs 2020). As a result, residents have a huge impact on the future of the city, and the decisions it makes. Cities are constantly being transformed by the decisions of democratically elected officials and the people (Natalie Bicknell Argerious 2020).

Generally speaking, residents control the direction of their city. Developments, Costs, Facilities, and more are done based on the demographic of the city, and what the residents want. To obtain this data, cities have a variety of collection methods such as surveys, and annual census'. Due to the amount of data greatly increasing in recent years, statistical analysis has never been more important, and analyzing resident data offers insights to guide the city in choosing it's next steps.

The remainder of this paper analyzes and discusses the results that were formed from survey data collected by the city of Toronto ("City of Toronto Open Data" 2009). The following sections will show the statistical methodologies used in this analysis, which all adhere to the

*Code and data are available at: https://github.com/Niyer02/Toronto-Public-Engagement-Analysis

best data science practices. This analysis was performed in R(R Core Team 2022), using tidyverse(Wickham et al. 2019), tibble(Müller and Wickham 2023), dplyr(Wickham et al. 2023), ggplot2(Wickham 2016), viridis(Garnier et al. 2023), knitr(Xie 2023), and KableExtra(Zhu et al. 2024). The desired outcome is that the reader fully understands the methods used, and how the end result was derived.

2 Data

The raw data was collected from Open Data Toronto ("City of Toronto Open Data" 2009), and was cleaned in R (R Core Team 2022) using the tidyverse (Wickham et al. 2019) and dplyr (Wickham et al. 2023) packages. The raw data consisted of 2 types or variables. The first being engagement questions, with binary response values `(NA, 1)`. The second being perception questions with responses ranging from (Strongly agree - Strongly disagree).

Feature engineering is the process of converting raw data (Or in our case, cleaned data) into use able features. We performed feature engineering on the cleaned data set to end up with our desired features, which can be seen in Table 1:

Table 1: Base Cleaned Data

engagement_prop	sentiment_average
0.25	0.45
0.55	0.50
0.10	0.20
0.65	-0.40
0.25	-1.20
0.25	0.15

2.1 Variables

The first variable is `engagement_prop` (Engagement Proportion). There were a total of 21 engagement questions, each associated with a distinct type of engagement (Survey, Consultation, etc.), this variable is the mean of all those questions. It is important to note that all engagement questions were structured such that `1` is positive (Participant engaged with the city via this method) and `0` is negative (Participant did not engage with the city via this method). Therefore, higher values of this variable can be interpreted as higher engagement with the city by the participant, while lower values can be interpreted as lower engagement with the city by the participant.

The second variable is `sentiment_average` (Sentiment Average). There were a total of 22 perception questions. Each question was structured s.t `2` is a strong positive sentiment, `-2` is

a strong negative sentiment, `0` is neutral, and `1` and `-1` represent a lesser positive and negative sentiment respectively and this variable is the mean of all those questions. Thus, values in the range of `[2, 0)` can be interpreted as positive sentiment, and values in the range of `(0, -2]` can be interpreted as negative sentiment, with the magnitude being represented in the numeric scale.

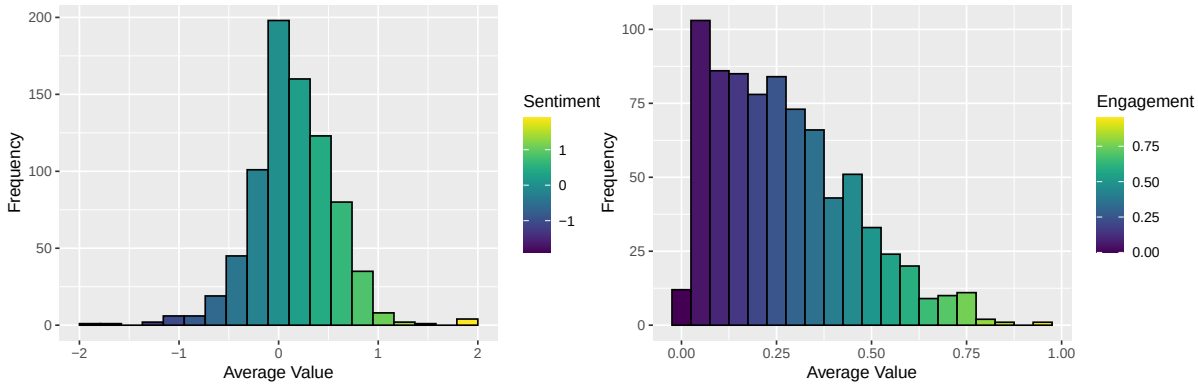
Table 2: Summary statistics for `sentiment_average`

Variable	n	Mean	SD	Median	Min	Max	Skew	Kurtosis
sentiment_average	792	0.150305	0.4159707	0.15	-2	2	-0.0347103	2.910171

Table 3: Summary statistics for `engagement_prop`

Variable	n	Mean	SD	Median	Min	Max	Skew	Kurtosis
sentiment_average	792	0.2707702	0.1802542	0.25	0	0.95	0.7326325	0.0255641

From Table 2 we can see that the mean sentiment is slightly positive, and has a standard deviation (SD) of 0.41. The skew of the graph is very slightly towards a negative sentiment. The distribution has a kurtosis of 2.9 which implies it has a high peak, and little tails. From Table 3 we can see that the mean engagement is 0.27, which is an average of 5.6 distinct methods of engagement with the city. The distribution also has a standard deviation (SD) of 0.18 The skew and kurtosis are not worth mentioning, as the distribution is modeled on binary data. The distributions of the variables can be seen in Figure 1. See Figure 1a for `sentiment_average` and Figure 1b for `engagement_prop`.



(a) Distribution of `sentiment_average`

(b) Distribution of `engagement_prop`

Figure 1: Distributions for variables of interest

3 Results

The scatter plot of `engagement_prop` and `sentiment_average` can be seen in Figure 2. The expected behavior of the trend line is a positive growth as engagement increases, however as seen in Figure 2, the trend line stays flat, and even decreases near the end. This shows a very low correlation between `engagement_prop` and `sentiment_average`. In this context, a low correlation between the two variables means that as a participant engages more and gives more data to their city, their overall sentiment towards the city remains the same, or slightly decreases. The inference drawn from this analysis is that Toronto may not prioritize its residents' opinions as much as it should, which leads to the residents' sentiment towards Toronto slightly decreasing.

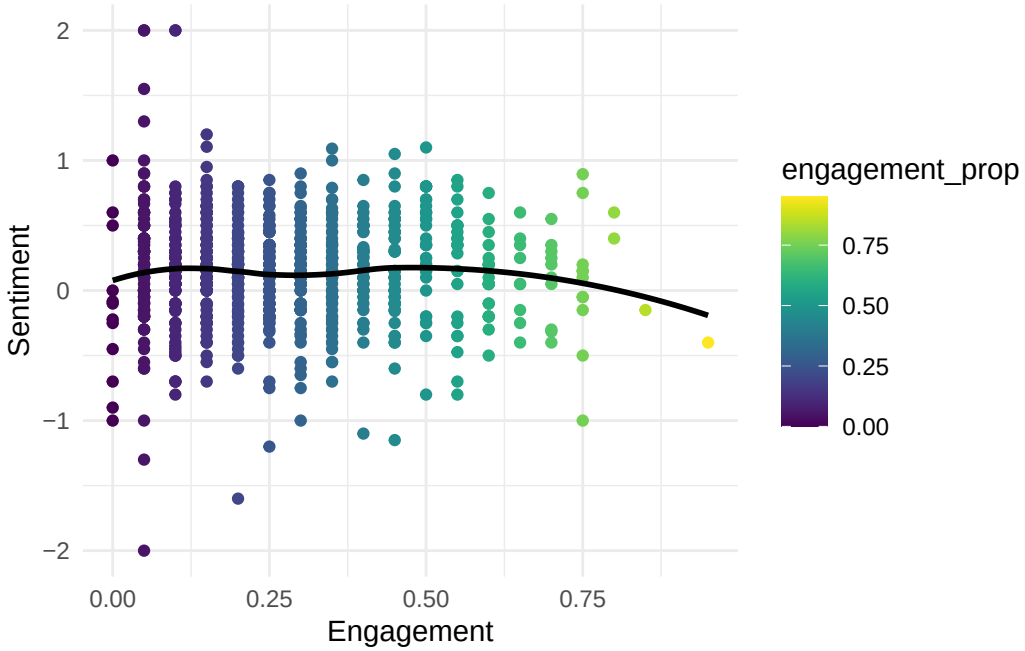


Figure 2: Scatterplot of `engagement_prop` and `sentiment_average`

4 Discussion

4.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

4.2 Second discussion point

4.3 Third discussion point

4.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected by, the data

Figure 3: **?(caption)**

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algorithm

Figure 4: **?(caption)**

References

"City of Toronto Open Data." 2009. https://open.toronto.ca/dataset/public-engagement-review-survey/.

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, et al. 2023. *viridis(Lite) - Colorblind-Friendly Color Maps for r*. https://doi.org/10.5281/zenodo.4679423.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*.

Natalie Bicknell Argerious. 2020. "Democracy and Cities." https://www.theurbanist.org/2020/11/03/democracy-and-cities/.

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

United Nations Department of Economic and Social Affairs. 2020. "Urbanization: Expanding Opportunities, but Deeper Divides." https://www.un.org/development/desa/en/news/social/urbanization-expanding-opportunities-but-deeper-divides.html.

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, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao, Will Beasley Thomas Travison and Timothy Tsai and, Yihui Xie, Rob Shepherd GuangChuang Yu and Stéphane Laurent and, Yoni Sidi, Brian Salzer, George Gui, et al. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra .