Is There a Correlation Between Different Licensed Business Categories That Share One Postal Code?[*]

An Analysis from the 2022 Toronto Business Licenses

Allyson Cui

18 September 2023

Abstract

The clustering of different types of businesses in a localized area can provide insights into economic synergies and potential areas for policy intervention. Our study utilizes Toronto OpenData for the year 2022 to examine the relationship between 68 business categories. The findings may inform local planning decisions and offer a quantitative methodology for studying business clusters.

Contents

1 Introduction 1

2 Data 2
- 2.1 Data Sources 2
- 2.2 Data Selection and Preparation 2
- 2.3 Data Description 3

3 Exploratory Data Analysis 4
- 3.1 Figure 1: Categories with the Highest Maximum Number of Licenses in a Single Postal Code 4
- 3.2 Figure 2: Categories with the Highest Average Number of Licenses 5
- 3.3 Figure 3: Categories with the Highest Total Number of Licenses 6
- 3.4 Integrated Insights 6

4 Correlation Analysis 6
- 4.1 Methodology 7
- 4.2 Results and Interpretation 7
- 4.3 Statistical Summary 9

5 Discussion 9

Appendix 11

1 Introduction

Background: Geographical business clusters represent a fascinating aspect of modern economics, embodying a localized synergy among diverse business categories [Carbonara, 2005]. These clusters often emerge as hotbeds of innovation, enhanced production, and efficient resource allocation [Lublinski, 2002]. However, not

[*] Code and data are available at: https://open.toronto.ca/dataset/municipal-licensing-and-standards-business-licenses-and-permits/

all business types are inclined to cluster together, leading to a complex landscape where spatial proximity does not always indicate functional interaction or collaboration. Existing literature focuses on the identification of these clusters, particularly in specialized sectors like aeronautics [Lublinski, 2002]. Furthermore, other works delve into how these clusters evolve their innovation capabilities, based on a model that adapts to the cluster's development stage [Bittencourt et al., 2019]. While these studies offer valuable insights, there is an evident gap in the quantitative analysis of what business types tend to co-locate and the nature of their interactions within a generalized urban setting.

Literature Review: To fill this gap, the present study leverages OpenData for Toronto 2022 to undertake an exhaustive analysis of 81 prescribed business categories and their spatial relationships. By creating 3828 unique pairs of business category combinations, this research aims to scrutinize the extent to which certain categories display valid correlations in their geographic distribution. Unlike previous studies that focused primarily on specialized sectors or qualitative aspects [Lublinski, 2002, Bittencourt et al. [2019]], our research employs statistical methods to provide a comprehensive, data-driven perspective.

Research Gap: This paper is organized as follows: the first section reviews existing literature on geographical business clusters, highlighting the methodologies employed and gaps that need to be filled. The second section describes the methodology adopted in this study, followed by a section detailing the findings. Finally, the paper concludes with a discussion of the implications of our research, and suggestions for future studies in this domain.

Objectives and Stucture of This Paper: In summary, this study contributes to the existing body of research by introducing a quantitative methodology to examine the nature and extent of business clustering in a large urban area. By doing so, it offers a multi-faceted understanding of the phenomenon, adding a valuable layer of complexity to our current understanding of geographical business clusters. The study is organized as follows: Section 2 reviews the existing literature; Section 3 describes the methodology; Section 4 presents the data and findings; and the final section discusses the implications of the study and offers suggestions for future research.

2 Data

2.1 Data Sources

The pivotal variables for our research question encompass both the category of business and its geographical location. This raises questions around the definition of 'category' and the precision level of 'location' that could serve as a meaningful measure of clustering. Given that the study's focus is on licensed businesses, the most logical approach is to adopt the government's classification criteria used during the licensing process. For location, we had options ranging from city-wide data to postal codes and even street numbers. We ultimately chose postal codes as our geographical unit of interest. Cities are too broad to offer valuable insights into clustering, while street numbers may be overly specific. A single postal code typically spans an area of 0.5 to 1 square kilometer—a walkable distance for most people—and is therefore an appropriate scale for studying potential relationships between businesses within the same locale. Additionally, the temporal dimension is crucial; a business license generally has both an issue and expiration date, indicating the period during which businesses might cluster.

Our data source is the "Municipal Licensing and Standards - Business Licenses and Permits" dataset, collected by Toronto's Municipal Licensing & Standards (ML&S) and available on the Toronto Open Data Portal [Gelfand, 2020].

The project had made use of R [R Core Team, 2020], and R packages "tidyverse" [Wickham et al., 2019], "dplyr" [Wickham et al., 2021], "janitor" [Firke, 2021], "knitr" [Xie, 2021], "stringi" [Gagolewski, 2022], "readr" [Wickham et al., 2023], "lubridate" [Grolemund and Wickham, 2011], "purrr" [Wickham and Henry, 2023], "testthat" [Wickham, 2011], and "assertthat" [Wickham, 2017].

2.2 Data Selection and Preparation

Our initial dataset contained 155,037 observations across 18 variables. These variables included the license category, business address at the postal code level, date issued, and cancel date. Out of an initial 81 categories and 16,607 postal codes, we were able to distill the dataset down to 3,919 postal codes and 68 categories by removing irrelevant or insufficient data. This included outdated licenses, postal codes featuring only one type of business, and postal codes that were too broadly defined (only first three characters available). We also removed businesses that had postal codes marked as "NA."

Table 1: Number of Licence per Category by Postal Code: Sample Data

Postal Code	ADULT ENTERTAINMENT CLUB	BILLIARD HALL	BODY RUB PARLOUR
K0L 1L0	0	0	0
K0L 1T0	0	0	0
K0M 2B0	0	0	0
L0B 1A0	0	0	0
L0C 1A0	0	0	0

2.3 Data Description

Focusing on the most recent data available from the year 2022, our target dataset includes businesses with licenses that were active for any duration within that year. This means that the issue date of the license is no later than December 31, 2022, and the cancelation date is no earlier than January 1, 2022. While it's possible that the operating timelines for some businesses may not completely overlap, we consider their inclusion valuable for our analysis. We operate under the assumption that these businesses are likely to be active concurrently, given the close proximity of their operational timelines. Accordingly, you won't find a dedicated 'time' variable in our cleaned data table, referred to as Table 1.

After filtering out postal codes that didn't provide meaningful data for our study, we narrowed down the number of postal codes from the original 16,607 to 3,919. Subsequently, we also removed business categories not found in these 3,919 postal codes, reducing the number of categories from 81 to 68. We restructured the dataset to display the number of licenses per business category, organized by postal code. This information is presented in Table 1, where both rows and columns are sorted in ascending order.

It's worth noting that although the sample data in Table 1 doesn't display non-zero values, but it gives a clear representation of how the dataset is organized. You may also observe that some postal codes in the sample don't start with the letter 'M,' which typically designates areas within the Greater Toronto Area (GTA). This is because some businesses licensed by the Toronto municipal government are not headquartered within the GTA. We chose not to exclude these businesses from our analysis, as their inclusion does not adversely affect our study of category correlation. Businesses located in other regions can still contribute valuable data, as they too interact with and are influenced by neighboring businesses, some of which may also hold licenses from the Toronto municipal government.

3 Exploratory Data Analysis

Note that the data now is focusing on the postal codes which have at least two different business categories licensed.

3.1 Figure 1: Categories with the Highest Maximum Number of Licenses in a Single Postal Code

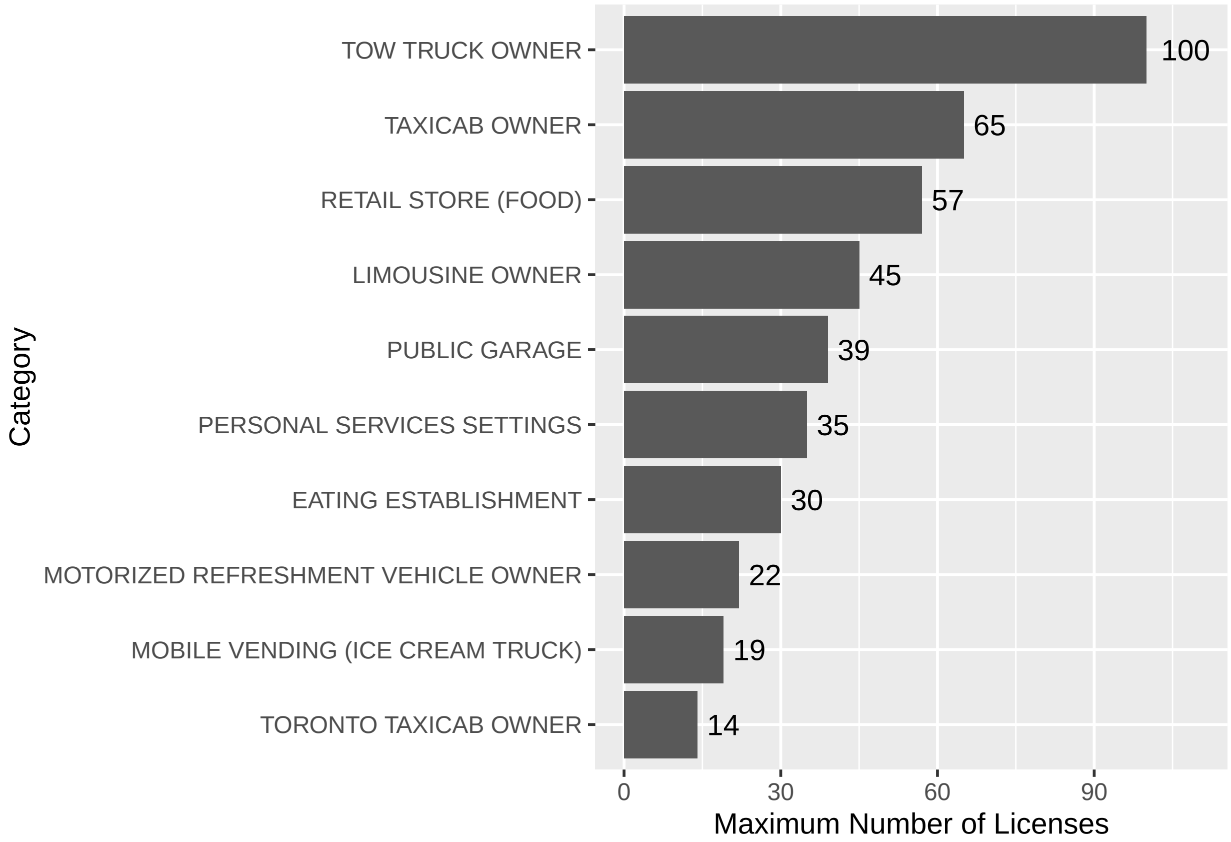

Figure 1: Top 10 Categories with the Highest Maximum Number of Licenses in a Single Postal Code

Figure 1 provides a view of business categories that have the highest concentration within individual postal codes. This can be seen as an indicator of highly localized monopolies or clusters. For example, "Tow Truck Owner" tops the chart with a maximum of 100 licenses in a single postal code, followed by "Taxicab Owner" with 65 and "Retail Store (Food)" with 57.

Businesses with Large Fleets: Businesses like tow truck and taxicab owners tend to cluster heavily in specific postal codes. This could imply the presence of depots or centralized locations for these services.

Consumer Essentials: Retail stores offering food items also show a significant maximum number of licenses in single postal codes, which might indicate areas of high residential density where convenience stores and supermarkets are plentiful.

Diverse Services: The list includes a variety of services from automotive ("Public Garage") to food and beverages ("Eating Establishment") and even specific types of vending ("Mobile Vending").

3.2 Figure 2: Categories with the Highest Average Number of Licenses

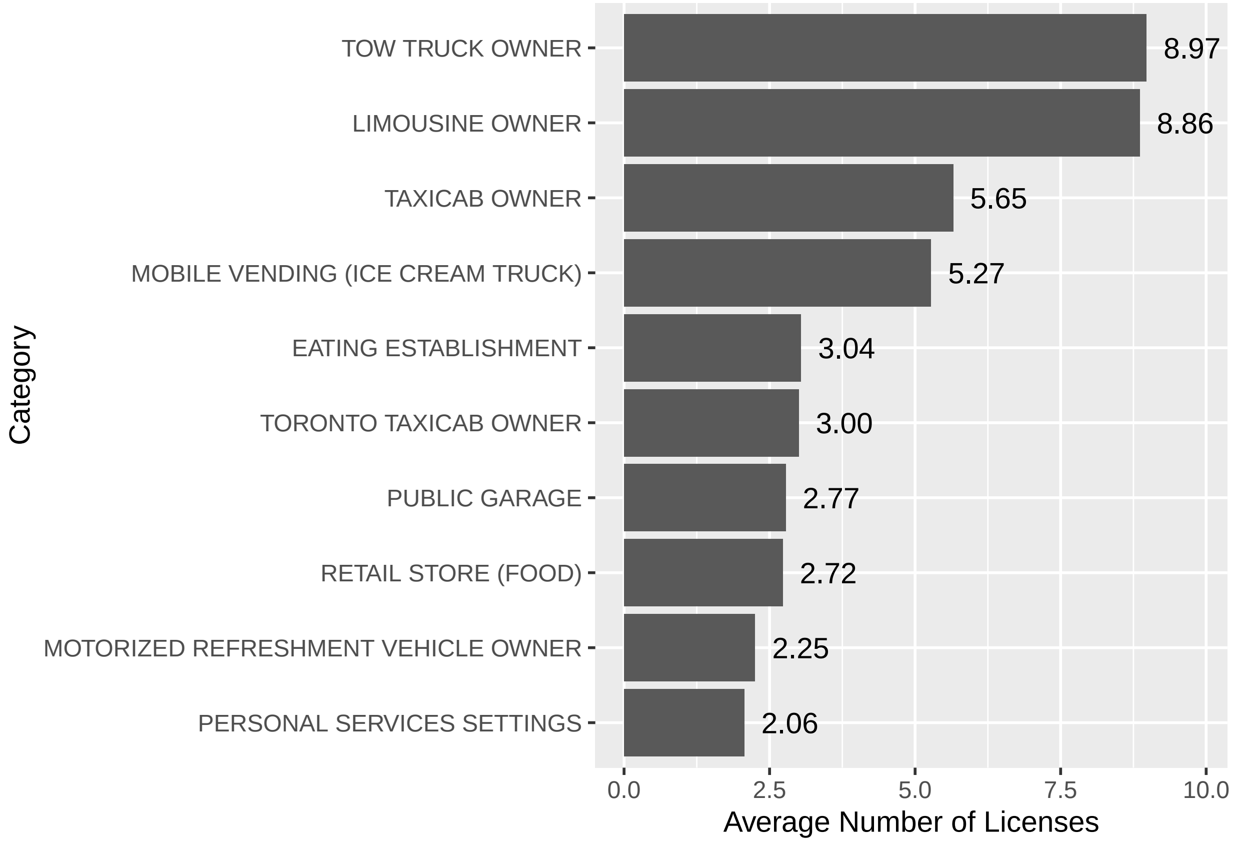

Figure 2: Top 10 Categories with the Highest Average Number of Licenses, Ignoring Postal Codes with No License of That Particular Category

Figure 2 moves from extreme concentration to provide a more balanced view by considering the average number of licenses across all postal codes, excluding those where a particular category is not represented.

Consistent Players: Categories like “Tow Truck Owner” and “Limousine Owner” appear again, but their numbers are now normalized, giving us a more generalized view of their prevalence.

Specialized Services: Categories such as “Mobile Vending (Ice Cream Truck)” make an appearance, showing they have a consistent but specialized presence across multiple areas.

Consumer Staples Remain: The average number of “Retail Store (Food)” and “Eating Establishment” licenses are still relatively high, further substantiating their widespread existence across the city.

3.3 Figure 3: Categories with the Highest Total Number of Licenses

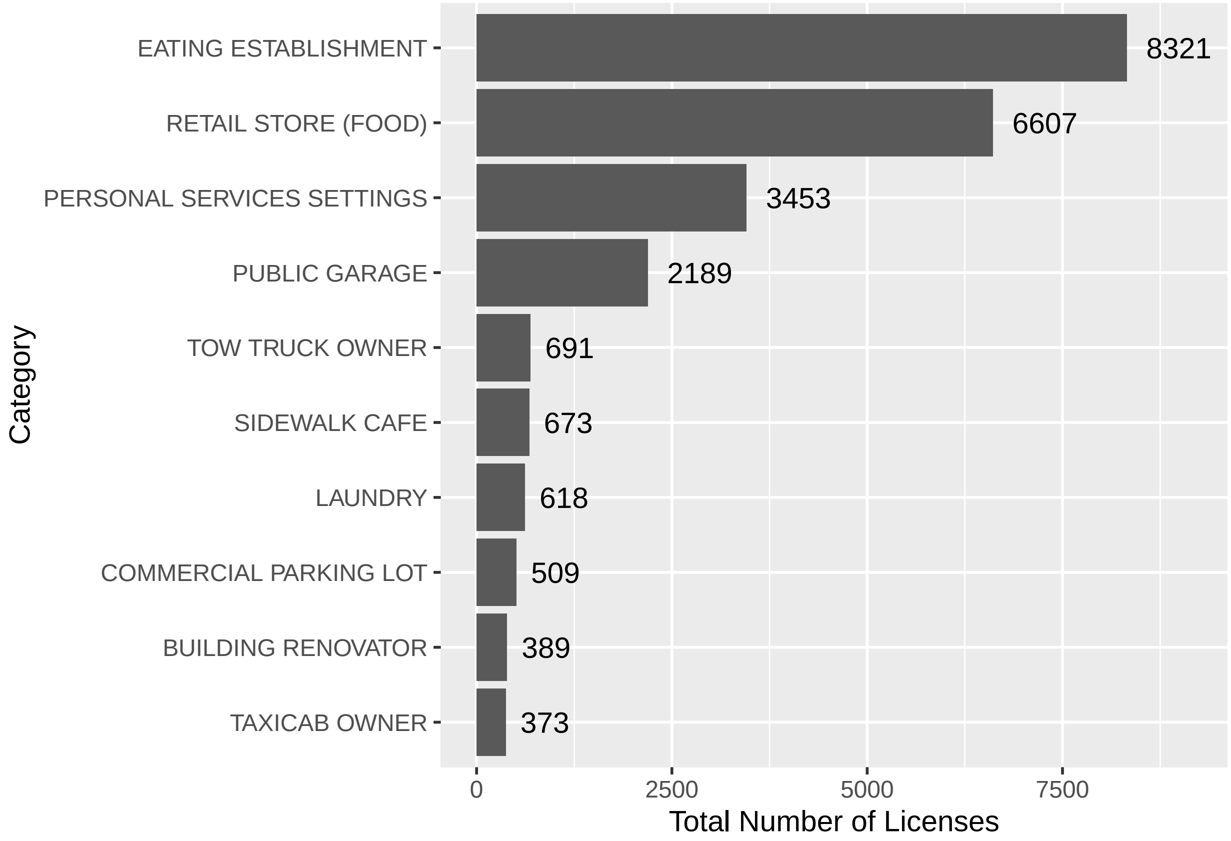

Figure 3: Top 10 Categories with the Highest Total Number of Licenses among All Postal Codes

Figure 3 gives an overall view by showing the total number of licenses across all postal codes for each category.

Broad Appeal: "Eating Establishment" and "Retail Store (Food)" lead the pack by a significant margin, suggesting that they are omnipresent and cater to a broad consumer base. **Essential Services**: Categories like "Personal Services Settings" and "Public Garage" indicate services that are not only essential but are also widespread. **Niche Categories**: Categories like "Sidewalk Café" and "Laundry" also make the list, showing that specialized businesses too have a notable presence.

3.4 Integrated Insights

Overlap and Complementarity: Businesses like eating establishments and retail food stores appear in all graphs, suggesting that they are not just popular but also well-distributed. The clustering of these business types might suggest a complementary relationship, which could be explored further.

Market Saturation: The high concentration of specific business types like "Tow Truck Owner" in some areas may indicate market saturation or potentially high levels of competition.

4 Correlation Analysis

The aim of this section is to examine the relationships between different categories of business licenses. Understanding these correlations can provide unique insights into business patterns, zoning considerations, and market dynamics.

4.1 Methodology

The correlation analysis begins with a cleaned dataset of business licenses across various categories and their postal code locations. We then compute a correlation matrix using built-in statistical functions. Unique pairwise combinations of business categories were extracted, and their respective correlation coefficients were calculated. This resulted in a table with three columns: Category A, Category B, and the Correlation Coefficient between them.

4.2 Results and Interpretation

4.2.1 Top 10 Highest Correlation Coefficients

Table 2: Top 10 Highest Correlation Coefficients

Category A	Category B	Correlation Coefficient
MOBILE VENDING (ICE CREAM TRUCK)	MOTORIZED REFRESHMENT VEHICLE OWNER	0.93
BUILDING CLEANER	LIMOUSINE OWNER	0.69
TAXICAB OPERATOR	TAXICAB OWNER	0.65
CARNIVAL	CIRCUS	0.50
LIMOUSINE OWNER	LIMOUSINE SERVICE COMPANY	0.47
EATING ESTABLISHMENT	RETAIL STORE (FOOD)	0.40
CURB LANE CAFE	SIDEWALK CAFE	0.38
PERSONAL SERVICES SETTINGS	RETAIL STORE (FOOD)	0.35
EATING ESTABLISHMENT	PERSONAL SERVICES SETTINGS	0.35
EATING ESTABLISHMENT	SIDEWALK CAFE	0.33

Table 2 shows the business categories with the highest correlation coefficients. Notably, the highest correlation is between “Mobile Vending (Ice Cream Truck)” and “Motorized Refreshment Vehicle Owner” at 0.93044.

The business categories that appear at the top with high positive correlation coefficients generally belong to related or complementary sectors. For example, “Mobile Vending” and “Motorized Refreshment Vehicle Owner” both serve food and refreshments in a mobile setting, likely appealing to similar customer bases and benefiting from being in proximity to each other.

The presence of other similar pairings, such as “Taxicab Operator” and “Taxicab Owner” or “Limousine Owner” and “Limousine Service Company,” further supports the idea that businesses in the same or complementary sectors tend to establish themselves in similar areas. This could be due to shared operational requirements like zoning laws, similar target demographics, or the advantages of clustering—for example, creating a “food truck hub” or “taxi stand” that attracts more customer attention than a solitary business could.

4.2.2 Top 10 Lowest Correlation Coefficients

Table 3: Top 10 Lowest Correlation Coefficients

Category A	Category B	Correlation Coefficient
COMMERCIAL PARKING LOT	PUBLIC GARAGE	-0.08
PERSONAL SERVICES SETTINGS	PLUMBING CONTRACTOR	-0.08
PUBLIC GARAGE	SIDEWALK CAFE	-0.08
BUILDING RENOVATOR	PERSONAL SERVICES SETTINGS	-0.08
PERSONAL SERVICES SETTINGS	PUBLIC GARAGE	-0.08
PLUMBING CONTRACTOR	RETAIL STORE (FOOD)	-0.09
COMMERCIAL PARKING LOT	PERSONAL SERVICES SETTINGS	-0.09
BUILDING RENOVATOR	RETAIL STORE (FOOD)	-0.10
EATING ESTABLISHMENT	PLUMBING CONTRACTOR	-0.12
BUILDING RENOVATOR	EATING ESTABLISHMENT	-0.14

Table 3 shows the business categories with the lowest correlation coefficients. Most notably, the most negative correlation exists between "Building Renovator" and "Eating Establishment," with a coefficient of -0.14225.

On the other end of the spectrum, the business categories with the most negative correlations typically have contrasting needs or target different consumer bases. For instance, "Building Renovators" might prefer areas that are under development or renovation and may not be operational during standard business hours, whereas "Eating Establishments" would likely want stable, high-traffic areas that are already developed.

However, it's interesting to note that these negative correlations are not as extreme as the positive ones. This suggests that while similar businesses have strong reasons to cluster together, dissimilar businesses don't necessarily avoid each other to the same extent—they may be indifferent to each other's presence rather than antagonistic.

4.2.3 Density Plot of Correlation Coefficients

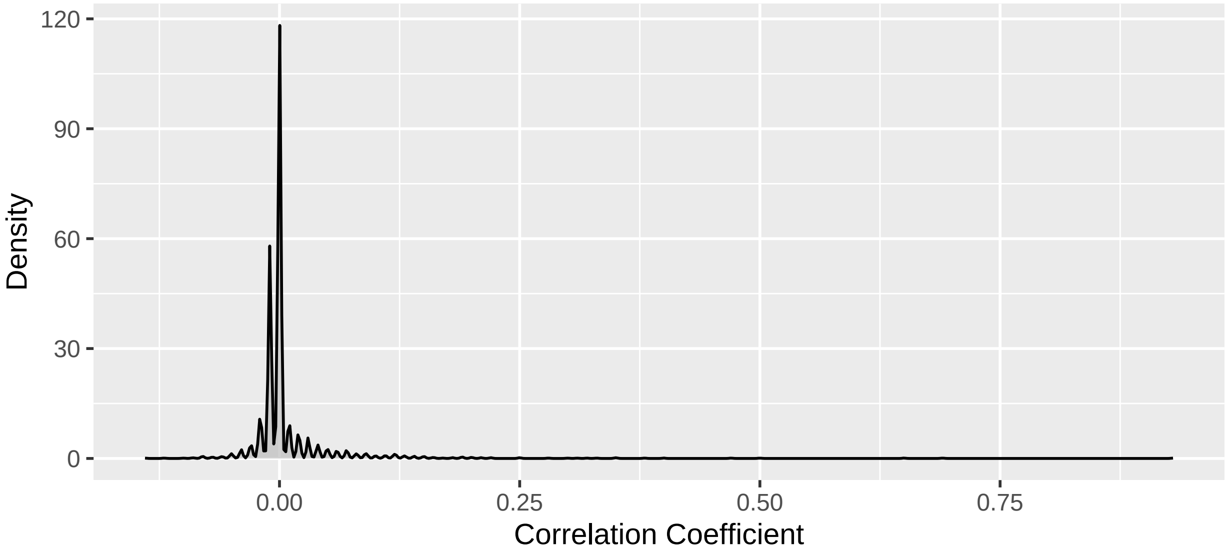

Figure 4: Density Plot of Correlation Coefficients

Figure 4 of correlation coefficients reveals a near-normal distribution, skewed slightly towards zero.

Table 4: Statistical Summary of Correlation Coefficients (Part 1)

...1	vars	n	mean	sd	median	trimmed	mad
X1	1	2278	0.0066533	0.0498584	-0.0026931	-0.0014627	0.0050608

Table 5: Statistical Summary of Correlation Coefficients (Part 2)

...1	min	max	range	skew	kurtosis	se
X1	-0.1422461	0.9304364	1.072682	7.610391	95.10072	0.0010446

The majority of correlation coefficients are near zero, indicating that most business categories don't have a strong inclination to co-locate or avoid each other. However, the extreme positive correlations are much more pronounced than the extreme negative ones.

This pronounced disparity could be attributed to several factors. For one, businesses with high positive correlations might have compelling synergistic reasons to cluster—for instance, shared supply chain requirements, pooled customer bases, or similar regulatory environments. On the other hand, businesses with negative correlations may not have as strong reasons to avoid each other, thus leading to more modest negative coefficients. For example, a high-end boutique and a discount store might not target the same customers but could co-exist in a large shopping area without directly affecting each other's business.

In summary, the businesses at the top of the positive correlation list seem to have concrete, often symbiotic reasons to co-locate, while those at the negative end may simply be avoiding each other due to different target demographics or business needs. However, the lack of extremely negative coefficients suggests that most businesses are more indifferent rather than antagonistic to dissimilar business types when choosing locations.

This enhanced third section aims to provide a more thorough understanding of the correlation coefficients and the underlying reasons for why certain business categories are more likely to be found together or apart.

4.3 Statistical Summary

```
## \vspace{5mm}
```

The mean correlation coefficient is approximately 0.007, close to zero. This implies that high correlations are generally rare. The 95th percentile has a coefficient of 0.075, further confirming that extreme correlations are unusual.

5 Discussion

One of the primary limitations of this study is the assumption that businesses are operational throughout the entire period of their license validity. While this simplifies the analysis, it may not accurately reflect the complexities of business operations. Another limitation arises from our focus on data from only the year 2022, ignoring the rich historical data that dates back to as early as January 1st, 1948. This temporal limitation precludes us from examining any chronological or causal relationships among businesses over time. Furthermore, our data cleaning process led to the exclusion of potentially insightful variables, such as restrictions on the license or permit, the names under which companies operate, and more granular location data at the street number level. Additionally, using postal codes as geographical units could either dilute or amplify the correlation effect between certain business categories [Grubesic, 2008].

Future studies could significantly extend the scope and applicability of this work in several ways. First, a time-series analysis could be conducted to understand how business clusters evolve over time. This could provide insights into whether older businesses tend to attract newer ones within the same or complementary categories. Additionally, future research could incorporate variables that were excluded in this study, such as license restrictions and street-level location data, to provide a more nuanced understanding of business clustering. Moreover, understanding the role of temporal factors, like seasonality or economic cycles, could shed light on the dynamic nature of business clusters. Longitudinal studies that track these variables over time could offer invaluable insights into the external factors that influence the phenomenon of business clustering.

In this study, we delved into the fascinating phenomenon of business clustering within the context of Toronto's licensed businesses. Using data sourced from Toronto's Open Data Portal, specifically from "Municipal Licensing and Standards - Business Licenses and Permits," we analyzed the spatial distribution of various business categories at the postal code level. Our analysis uncovered that out of 3003 unique pairs of business categories, about 1% exhibited a correlation coefficient higher than 0.194, with the highest being a remarkable 0.930. This suggests that certain types of businesses such as food-related professions tend to co-locate, possibly benefiting from shared resources, customer bases, or other synergistic effects.

However, it's important to acknowledge the limitations of this research, such as the assumption that businesses operate throughout the entire validity period of their licenses, and the exclusion of potentially significant variables.

While these findings offer a starting point, much work remains to be done. Future research could enrich these insights by considering temporal factors, economic cycles, and other external conditions that weren't addressed in this study. Overall, the research serves as a preliminary step in understanding the complex dynamics of business clustering, and we hope it sparks further investigation into this economically and socially significant phenomenon.

Appendix

You may go to the ‘outputs>data>correlation_matrix.csv’ to review the complete list of combinations and their correlation coefficients.

References

Bruno Anicet Bittencourt, Aurora Carneiro Zen, and Frédéric Prévot. Innovation capability of clusters: understanding the innovation of geographic business networks. *REVIEW OF BUSINESS MANAGEMENT*, 21(4), 2019. ISSN 1806-4892. doi: 10.7819/rbgn.v21i4.4016.

Nunzia Carbonara. Information and communication technology and geographical clusters: opportunities and spread. *Technovation*, 25(3):213–222, 2005. ISSN 0166-4972. doi: 10.1016/S0166-4972(03)00095-6. URL https://www.sciencedirect.com/science/article/pii/S0166497203000956.

Sam Firke. *janitor: Simple Tools for Examining and Cleaning Dirty Data*, 2021. URL https://github.com/sfirke/janitor. R package version 2.1.0.

Marek Gagolewski. stringi: Fast and portable character string processing in R. *Journal of Statistical Software*, 103(2):1–59, 2022. doi: 10.18637/jss.v103.i02.

Sharla Gelfand. *opendatatoronto: Access the City of Toronto Open Data Portal*, 2020. URL https://sharlagelfand.github.io/opendatatoronto/,https://github.com/sharlagelfand/opendatatoronto/.

Garrett Grolemund and Hadley Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011. URL https://www.jstatsoft.org/v40/i03/.

Tony H. Grubesic. Zip codes and spatial analysis: Problems and prospects. *Socio-Economic Planning Sciences*, 42(2):129–149, 2008. ISSN 0038-0121. doi: 10.1016/j.seps.2006.09.001. URL https://www.sciencedirect.com/science/article/pii/S0038012106000516.

Alf Erko Lublinski. *Geographical business clusters: concepts for cluster identification with an application to an alleged aeronautics cluster in Northern Germany.* PhD thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2002.

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

Hadley Wickham. testthat: Get started with testing. *The R Journal*, 3:5–10, 2011. URL https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

Hadley Wickham. *assertthat: Easy Pre and Post Assertions*, 2017. URL https://CRAN.R-project.org/package=assertthat. R package version 0.2.0.

Hadley Wickham and Lionel Henry. *purrr: Functional Programming Tools*, 2023. https://purrr.tidyverse.org/, https://github.com/tidyverse/purrr.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL https://dplyr.tidyverse.org,https://github.com/tidyverse/dplyr.

Hadley Wickham, Jim Hester, and Jennifer Bryan. *readr: Read Rectangular Text Data*, 2023. https://readr.tidyverse.org, https://github.com/tidyverse/readr.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2021. URL https://yihui.org/knitr/. R package version 1.37.