Understanding Factors That Affects the Existence of Digital Innovation Hubs*

An Analysis of Toronto Libraries' Data

Yanzun Jiang

November 30, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/Stary54264/relationship_between_attributes_and_whether_dih_exists_in_toronto_libraries.

1 Introduction

Despite the growing importance of innovation hubs in the global economy, the factors influencing the presence of Digital Innovation Hubs (DIH) in facilities remain under-explored and needed to be filled. This paper seeks to identify the key factors that influence whether a library in Toronto has a DIH, focusing on attributes such as area size, the number of public parking spots, the number of workstations in the library, and the number of years from the library was built.

To investigate these factors, a logistic regression model was utilized, which is ideal for predicting binary outcomes (the presence or absence of a DIH). This analysis aims to estimate the likelihood of a library having a DIH based on these attributes. The dataset used in this study contains information on various libraries in Toronto. After data cleaning, relevant variables was selected to do data analysis. The logistic regression model will reveal how each attribute (area size, parking availability, number of workstations, and age of the building) impacts the presence of a DIH.

Results

DIHs are vital in fostering entrepreneurship, driving economic development, and supporting technological advancements. Understanding the attributes that contribute to the likelihood of a library of having a DIH could provide valuable insights for urban planners and policymakers.

The remainder of this paper is structured as follows: Section 2 discusses the data used in this analysis, including an overview of the dataset and key descriptive statistics. Section 3 outlines the logistic regression model, its assumptions, and its interpretation. Section 4 presents the results of the model, including the significance of each attribute in predicting the presence of a DIH. Section 5 offers a discussion of the findings, examines how these attributes affect DIH presence, and lists some weakness of this analysis. **?@sec-appendix** provides extra tables and graphs to further explain the findings in this paper.

Statistical analysis is performed using `R` (R Core Team 2023), with packages `tidyverse` (Wickham et al. 2019), `arrow` (Richardson et al. 2024), `janitor` (Firke 2023), `testthat` (Wickham 2011), `here` (Müller 2020), `modelsummary` (Arel-Bundock 2022), `performance` (Lüdecke et al. 2021), `knitr` (Xie 2014), `kableExtra` (Zhu 2024), and `rstanarm` (Brilleman et al. 2018).

1.1 Estimand

The estimand in this paper is the presence of a Digital Innovation Hub (DIH) in a library. However, accurately observing which libraries have a DIH is not straightforward due to several challenges faced by urban planners. They may not have access to complete or up-to-date information about all libraries. This lack of comprehensive data, combined with regional variations and different reporting standards, makes it difficult to directly know the presence

of a DIH across all libraries. To address this issue, a logistic regression model was built, since it allows us to estimate the probability of a facility having a DIH based on key attributes available in the dataset. By doing so, we aim to provide urban planners with insights into which factors most strongly influence the presence of DIHs, despite the challenges in directly measuring this estimand.

2 Data

This report uses a dataset collected by Toronto Public Library (Toronto Public Library 2024) and posted on Open Data Toronto (Gelfand 2022). Table 2 is a preview of the dataset. These data provide valuable insights into the physical attributes and temporal characteristics of libraries that may house DIHs. Alternative datasets, such as those focusing on municipal buildings or educational institutions, were considered but ultimately not used because libraries are more accessible to public, thus having a bigger impact on the society. The dataset includes predictor variables including library area, public parking spots, the number of workstations, and the year the library was built. Data on these variables would be used to predict the outcome variable - the presence of DIHs.

2.1 Variables

One outcome variable and four predictor variables are used in the model:

`dih`: The presence of DIHs in libraries in Toronto. If there is one or more DIHs in the library, this field is set to 1; otherwise, this field is set to 0.

`area`: The total size of the library measured in square feet.

`parking`: The number of parking spaces available for the public. If a branch does not have any public parking spaces or shares parking spaces with another location, this field is set to 0.

`workstations`: A count of computers with internet access available for public use in the branch.

`year`: The number of years from the year that the present location of the library was officially opened to the general public. This variable was constructed using the formula $year = 2024 - year\ built$. The age of the library was used instead of the year that the library was built since the age shows the degree of newness, which might affect the presence of DIHs. However, the year built is not that direct.

Their descriptive statistics and graphs of their distribution is included in Section A.

2.2 Measurement

The library dataset used in this study was collected through a combination of administrative records and self-reported surveys. While the dataset provides valuable insights, several factors need to be considered:

Reporting Variability: Libraries differ in how they track and report operational data, leading to potential inconsistencies. For instance, "workstations" might only include computers in some libraries, while others might also count other types of technological infrastructure.

Survey Limitations: The presence of a DIH was self-reported, which could result in over- or under-representation of actual DIHs. Some respondents might interpret "Digital Innovation Hub" differently based on their regional or institutional context.

Geographical Coverage: The dataset might not represent all regions equally. Libraries in rural areas may be underrepresented, impacting the generalization of the findings.

2.2.1 Connection with Real-World

The dataset connects real-world phenomena to measurable entries. For example:

`area`: The size of a library, measured in square feet, reflects its physical capacity to host community activities, including DIHs. Size of library would affect the presence of DIHs.

`parking`: The number of parking spaces captures the the likelihood of a library being visited. If a library is more likely to be visited, it would have more facilities.

`workstations`: Number of workstations symbolize a library's technological infrastructure - computers and other digital tools that enable public access to the internet, software, and other resources. Libraries with more computers is likely to have DIHs.

`year`: Newer libraries are often designed to meet current technological needs, making them more suitable for DIHs.

`dih`: The presence of a DIH at a library provides a learning opportunity for the public.

3 Model

In this analysis, a Bayesian logistic regression model was used to examine the relationship between the presence of a DIH in libraries and key library attributes.

Our assumptions include that samples could represent every library in Toronto, observations should be independent with each other, and no perfect multicollinearity between predictor variables.

There are some limitations of our model. The model would be no longer valid if the actual underlying relationship between the predictor variables and the outcome variable is non-linear. Also, the predictor variables might not follow normal distribution.

Additional predictor variables include position of the library (longitude and latitude). However, while the position of the library could affect the presence of DIHs indeed, the longitude and latitude are not associated with it linearly.

3.1 Model set-up

Define y_i as whether a library has a DIH (0 for no and 1 for yes) with probability p_i, so it follows a Bernoulli distribution. Then, the logistic link function of p_i is a linear combination of the intercept - β_0, and the effects of the predictor variables - β_1, β_2, β_3, and β_4 times the predictor variables respectively.

$$
\begin{aligned}
y_i|p_i &\sim \text{Bern}(p_i) && (1)\\
\text{logit}(p_i) &= \beta_0 + \beta_1 \times \text{area}_i + \beta_2 \times \text{parking}_i + \beta_3 \times \text{workstations}_i + \beta_4 \times \text{year}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)\\
\beta_3 &\sim \text{Normal}(0, 2.5) && (6)\\
\beta_4 &\sim \text{Normal}(0, 2.5) && (7)
\end{aligned}
$$

Since we have no prior information about any variable, the prior of the intercept and coefficients was set to a normal distribution with $\mu = 0$ and $\sigma = 2.5$. We run the model in R (R Core Team 2023) using the `rstanarm` package (Brilleman et al. 2018).

3.2 Model justification

We propose the following hypotheses regarding how library attributes influence the likelihood of having DIHs.

`area`: We anticipate that larger libraries are more likely to host DIHs. Greater area reflects increased physical capacity, enabling libraries to allocate dedicated spaces for innovation and technology.

`parking`: The number of parking spaces is expected to positively correlate with the presence of a DIH. Parking facilities make libraries more accessible to the public, encouraging diverse patronage and increasing community engagement. Libraries with better accessibility might

be more willing to improve its facilities and might receive greater funding and community support, promoting the establishment of advanced facilities like DIHs.

`workstations`: A strong positive relationship between the number of workstations and the presence of a DIH is expected. Workstations, representing technological infrastructure, are often central to the function of a DIH. Libraries with more workstations are likely equipped with the resources needed to support digital innovation and technological engagement.

`year`: We hypothesize a negative relationship between years since built and the likelihood of having a DIH. Older libraries might lack modern infrastructure and technological resources necessary for a DIH, whereas recently built or renovated libraries are more likely to own technological innovations aligned with the concept of a DIH.

4 Results

Our model results are summarized in Table 1. The findings align with our prior expectations in some degree and provide insights into the relationship between the predictor variables and the presence of DIHs. The intercept is negative, showing that more libraries do not have a DIH. The coefficient of `area` is 0, meaning the area of the library has litte effect on the outcome variable. The slope of `parking` and `workstations` are all positive, meaning a increase in them would lead to a higher likelihood of having DIHs. Rather, `year` has a negative slope, meaning older libraries are less likely to have a DIH, which aligns with our assumption.

The R^2 for this model is quite small, meaning the model might not fit the data well. However, the $RMSE$ is also quite small, meaning the problem of overfit would not present in our model. More details are included in Section B.

Table 1: Summary Statistics of the Logistic Regression Model

	(1)
(Intercept)	−1.915
area	0.000
parking	0.015
workstations	0.070
year	−0.063
Num.Obs.	100
R2	0.472
Log.Lik.	−15.332
ELPD	−20.6
ELPD s.e.	5.7
LOOIC	41.1
LOOIC s.e.	11.4
WAIC	40.4
RMSE	0.21

5 Discussion

5.1 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.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Descriptive Statistics and Graphs of the Dataset

A.1 Preview of the Dataset

Table 2: Preview of the Toronto Libraries Dataset

area	parking	dih	workstations	year
29000	59	1	38	7
28957	45	1	36	53
7341	0	0	7	25
27000	86	1	42	33
2988	0	0	5	42
7806	0	0	11	116

A.2 Descriptive Statistics

Table 3: Descriptive Statistics of `area`

Mean	Median	Minimum	Maximum	IQR
18129.3	8496.5	554	426535	8817.8

Table 4: Descriptive Statistics of `parking`

Mean	Median	Minimum	Maximum	IQR
12	0	0	139	15.2

Table 5: Descriptive Statistics of `workstations`

Mean	Median	Minimum	Maximum	IQR
17.9	11	2	213	12

Table 6: Descriptive Statistics of `year`

Mean	Median	Minimum	Maximum	IQR
53.8	49.5	1	117	28.2

A.3 Graphs

From Figure 1, we could see that more libraries do not have a DIH. All predictor variables are right-skewed (Figure 2, Figure 3, and Figure 4), except `year` (Figure 5). `year` might follows a normal distribution, while other predictor variables might follows an exponential distribution.

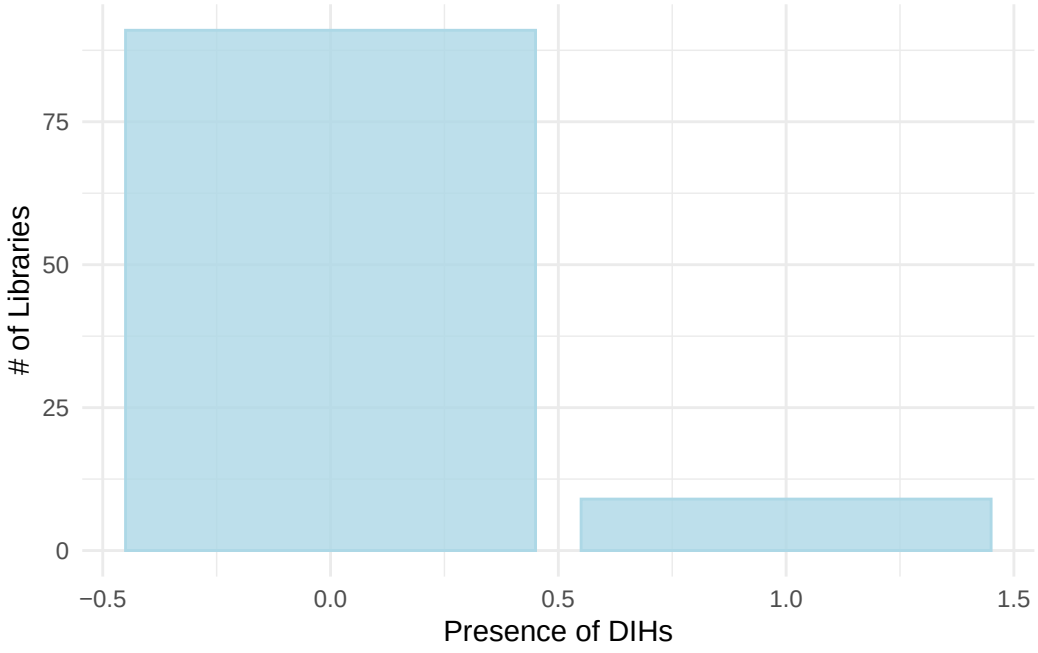

Figure 1: Presence of DIHs in Toronto Libraries

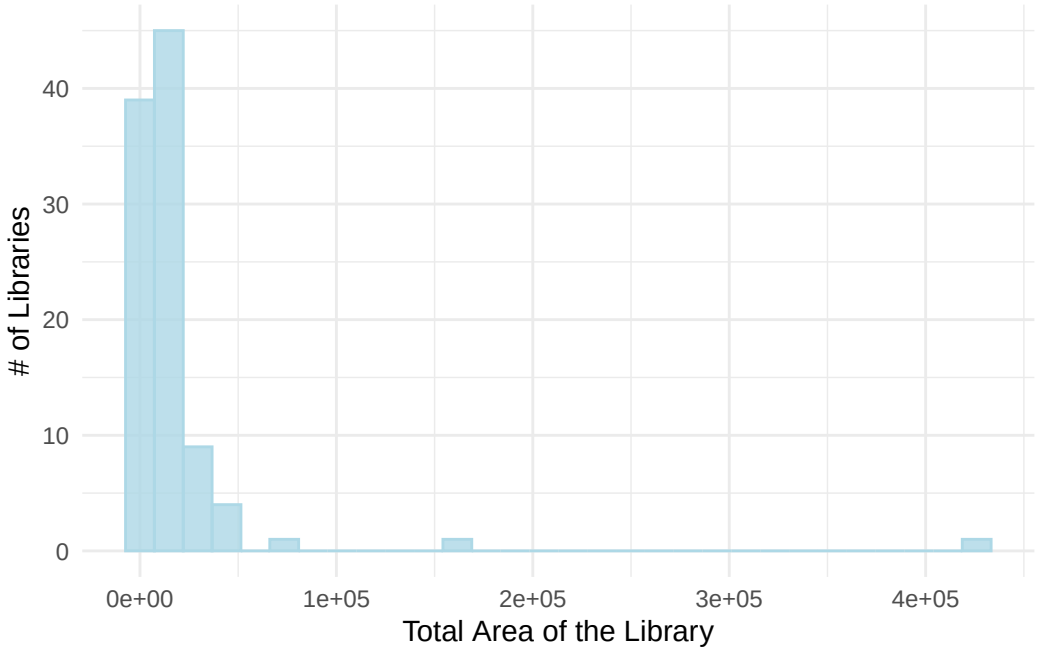

Figure 2: Distribution of Total Areas of Libraries in Toronto

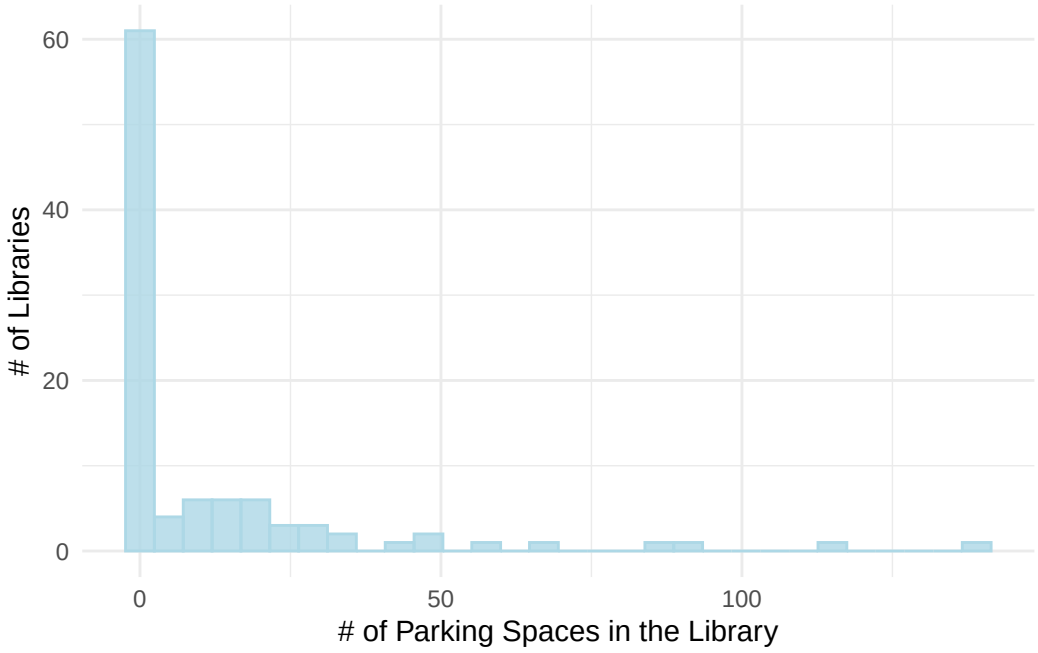

Figure 3: Distribution of Number of Parking Spaces at Libraries in Toronto

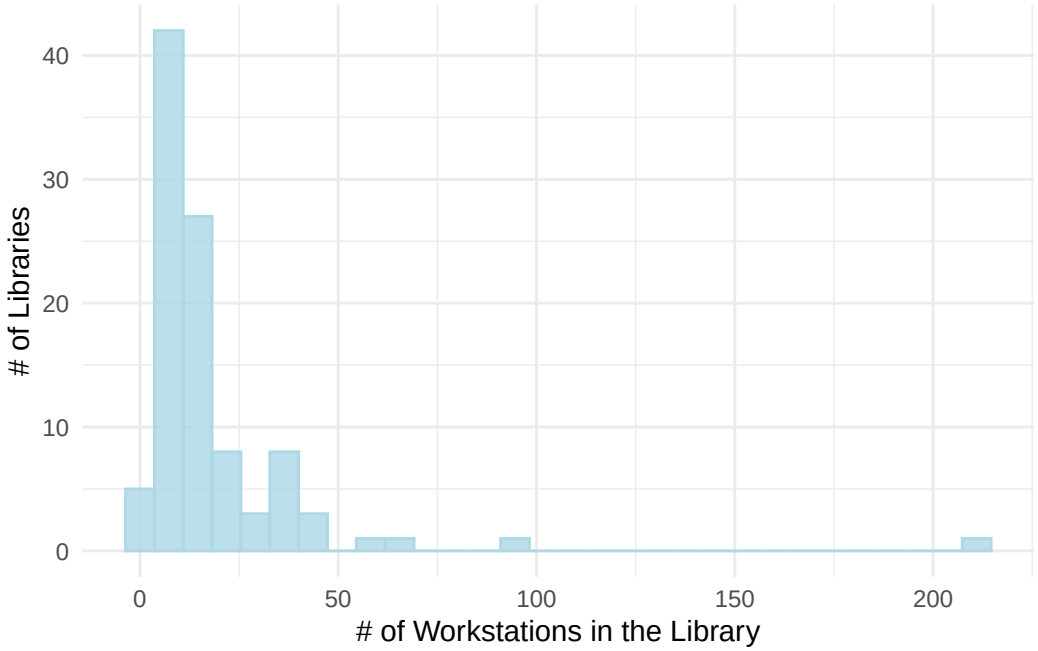

Figure 4: Distribution of Number of Workstations in Libraries in Toronto

Figure 5: Distribution of Age of Libraries in Toronto

B Model Details

B.1 Distribution of Posterior

The distribution of posterior is shown in Figure 6, which aligns with the actual data. From the graph, we can see that more libraries does not have a DIH.

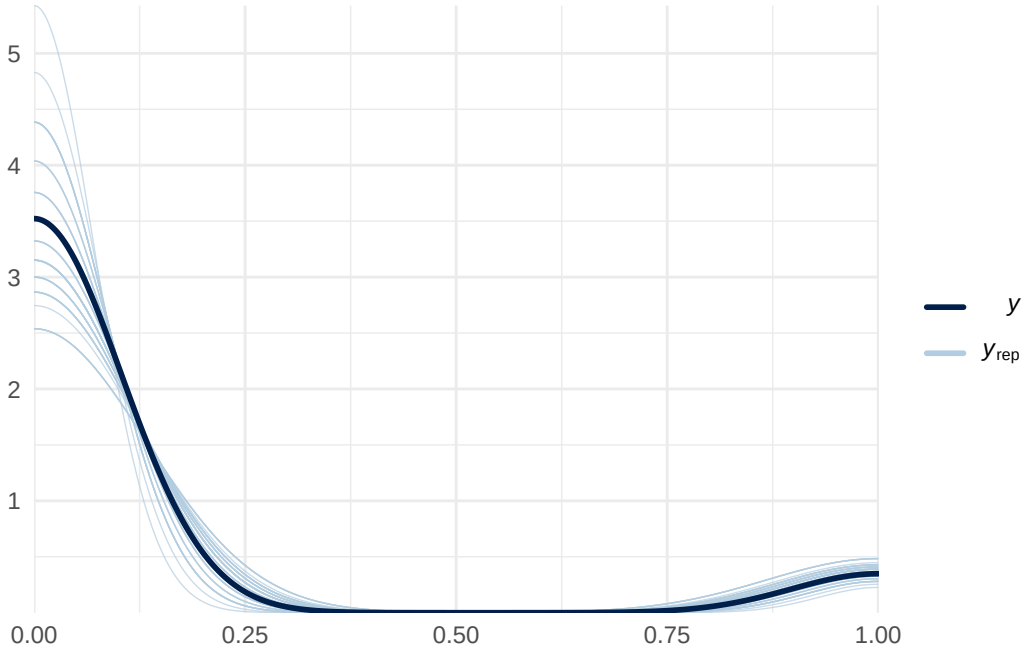

Figure 6: Posterior Distribution for the Logistic Regression Model

B.2 Density Plots

From Figure 7, we could see that smaller libraries are likely to have no DIHs, while larger libraries are not sure to have DIHs. In Figure 8 and Figure 9, libraries with more parking spaces and workstations tends to have a DIH. On the opposite, the pattern shown in Figure 10 tells us older libraries tend to have no DIH.

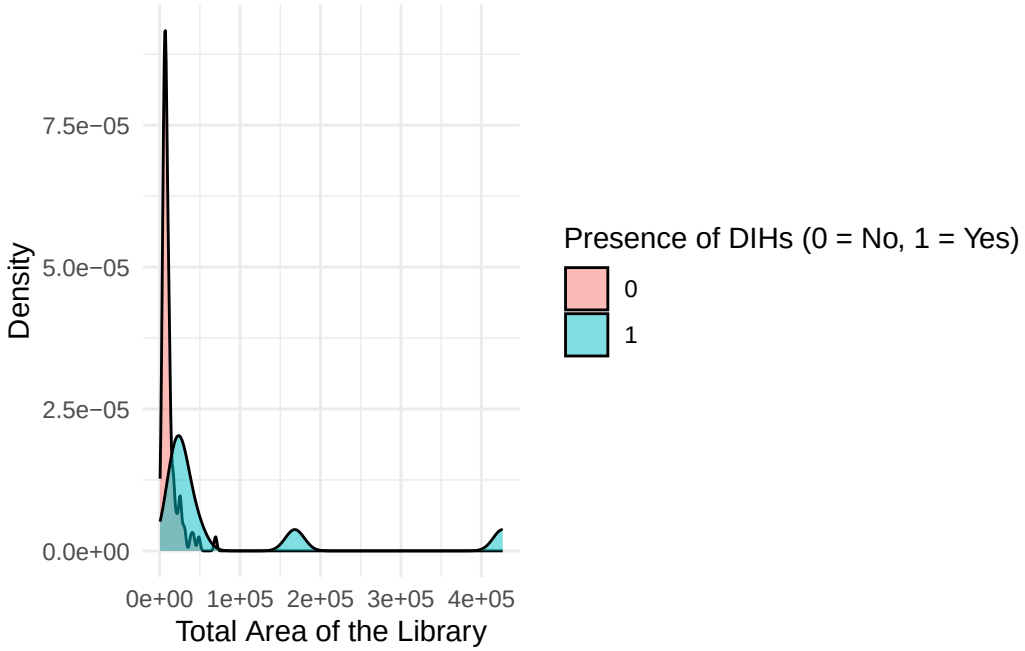

Figure 7

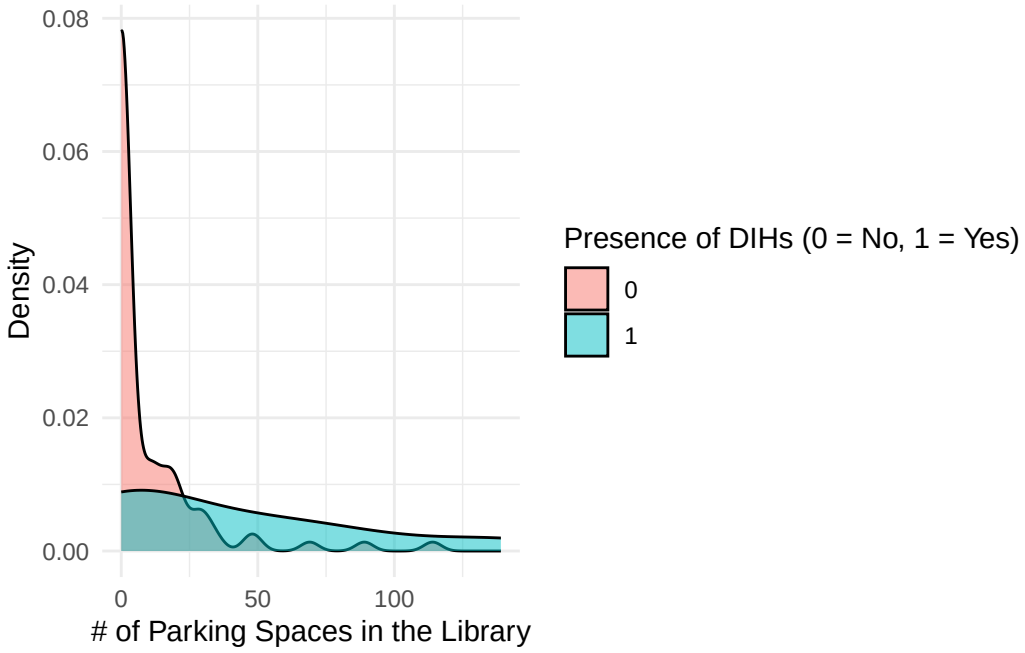

Figure 8

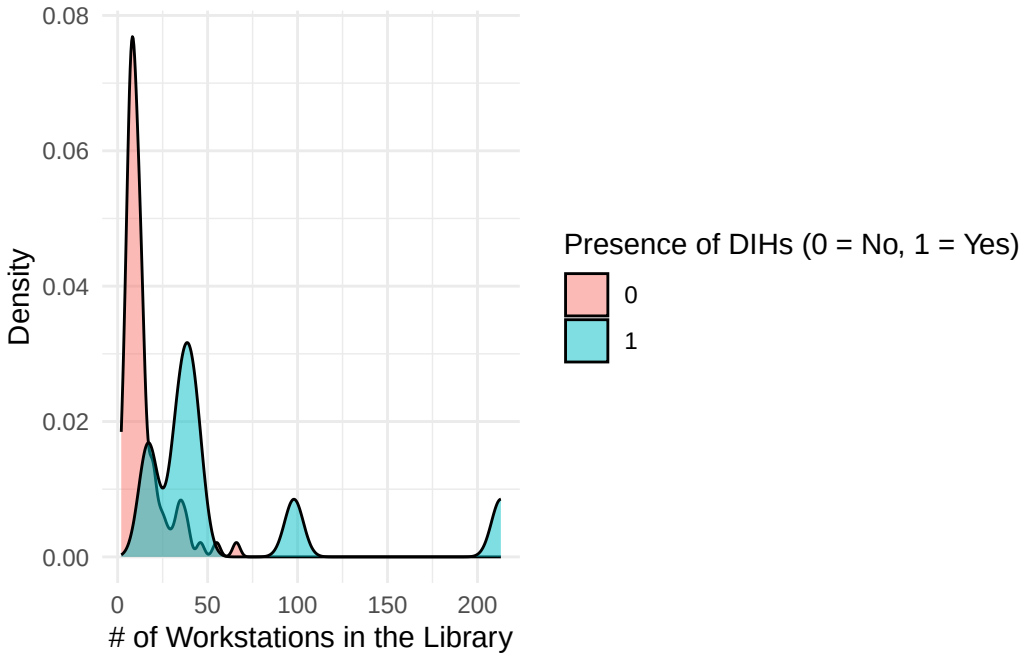

Figure 9

Figure 10

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Firke, Sam. 2023. *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.R-project.org/package=opendatatoronto.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. "performance: An R Package for Assessment, Comparison and Testing of Statistical Models." *Journal of Open Source Software* 6 (60): 3139. https://doi.org/10.21105/joss.03139.

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

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Toronto Public Library. 2024. "Library Branch General Information." https://open.toronto.ca/dataset/library-branch-general-information/.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

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.

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.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.