Understanding Factors That Affect the Existence of Digital Innovation Hubs*

An Analysis of Toronto Libraries' Data

Yanzun Jiang

December 2, 2024

This paper investigates what influences the presence of Digital Innovation Hubs (DIHs) in libraries, focusing on factors such as parking availability, workstations, library size, and age. Using a logistic regression model, we found that libraries with more parking and workstations are more likely to have DIHs, while older libraries are less likely to include them. These findings highlight the importance of modern facilities and adequate resources in fostering digital innovation within libraries. This research can guide library planners and policymakers in improving library services and ensuring access to technological opportunities for communities.

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 need to be filled. DIHs are vital in fostering entrepreneurship, driving economic development, and supporting technological advancements. Understanding the attributes contributing to the likelihood of a library having a DIH could provide clarity for urban planners and policymakers. 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 when the library was built.

A logistic regression model was utilized to investigate these factors, 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 were selected to do data analysis. The logistic regression model will show how each attribute (area size, parking availability, number of workstations, and age of the building) impacts the presence of a DIH.

In our analysis, the logistic regression model shows that libraries with higher parking and more workstations are more likely to have DIHs, while the area of the library has little effect on this outcome. Additionally, older libraries are less likely to feature a DIH, aligning with our expectations. The model's R^2 value suggests it does not fit the data very well, though the relatively low $RMSE$ indicates that overfitting is not a concern. These results provide clarity into the factors that influence the presence of DIHs in libraries.

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 limitations of this analysis.

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 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 clarity 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 the 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 clarity 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 the public, thus having a bigger impact on 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 are included in Section A.

2.2 Measurement

The library dataset used in this study was collected by Open Data Toronto (Gelfand 2022). While the dataset provides clarity, 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.

Geographical Coverage: The dataset might not represent all regions equally. Libraries in rural areas may be underrepresented since their information might be inaccurate and hard to access, 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. The size of the library would affect the presence of DIHs.

`parking`: The number of parking spaces captures the likelihood of a library being visited, and this would affect the facilities in it.

`workstations`: The number of workstations symbolizes a library's technological infrastructure - computers and other digital tools that enable public access to the internet, software, and other resources

`year`: Newer libraries are often designed to meet current technological needs.

`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 of 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 the normal distribution.

Additional predictor variables include the 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.

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

Since we have no prior information about any variable, the prior of the intercept and coefficients were 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. The 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 their facilities and might receive greater funding and community support, promoting the establishment of 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 the 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 to some degree and provide clarity 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 little effect on the outcome variable. The slope of `parking` and `workstations` are all positive, meaning an 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 overfitting would not be 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 Review of the Analysis

This paper investigates the factors influencing the presence of Digital Innovation Hubs (DIHs) in libraries. Libraries have evolved from being repositories of books to centers for community engagement, innovation, and technology access. DIHs are a relatively recent development in this transition, providing patrons with access to cutting-edge tools such as 3D printers, virtual reality systems, and other digital resources. However, the availability of DIHs is unevenly distributed, raising questions about what drives their presence in certain libraries but not others. Understanding these factors is important for policymakers, library administrators, and community planners aiming to promote equitable access to digital innovation.

To address this question, we used a logistic regression model to analyze library characteristics and their relationship with the likelihood of hosting a DIH. Specifically, we examined variables such as the size of the library (`area`), parking availability (`parking`), the number of workstations (`workstations`), and the library's age (`year`). These variables were chosen based on their potential relevance to the library's capacity to support innovation and accessibility.

Our analysis involved fitting a Bayesian logistic regression model, conducting diagnostic checks, and evaluating model performance using metrics such as R^2 and $RMSE$. This allowed us to identify significant relationships between library features and the presence of DIHs while accounting for potential model limitations. We supplemented the quantitative results with

graphical analyses to visualize the effects of key variables, enhancing the interpretability of our findings.

By focusing on these specific predictors, this paper provides clarity into actionable factors that can inform decision-making for libraries considering establishing DIHs. Additionally, the methods and findings provide a framework for future research exploring the integration of technology in community spaces. This analysis offers a perspective on how infrastructure and resource allocation shape technological accessibility, highlighting both strengths and gaps in existing library systems.

5.2 Physical Factors that Affect DIHs

One key outcome from this paper is the understanding of how library characteristics influence the likelihood of having a DIH. The logistic regression model shows that certain features - specifically parking availability, the number of workstations, and the library's age - are significant predictors of whether a library will have a DIH. This suggests that libraries with more parking and a higher number of workstations are more likely to offer DIHs.

From a broader perspective, this result shows how infrastructure and spatial resources can shape technological access. The availability of physical space for workstations and parking, for example, is not only a matter of practicality but may reflect deeper socio-economic factors, such as local investments in public services or the socio-economic status of the surrounding community. In particular, libraries located in areas with better parking availability and more resources may be better positioned to support digital hubs, thereby perpetuating a cycle of technological access in certain communities while excluding others.

This has far-reaching implications for policymakers and community planners. It suggests that resource allocation in libraries—especially for digital services—should be carefully considered, as certain library features can serve as enablers or barriers to offering DIHs. For example, if policymakers wish to increase the availability of DIHs in underserved areas, they may need to address not only the availability of technology and funding but also the structural features of libraries that make them suitable for innovation.

5.3 Relationship between Newness and DIHs

Another outcome from this paper is the relationship between a library's age and its ability to adapt to modern technological demands. The negative relationship between the age of a library and the presence of a DIH shows that older libraries are less likely to adopt or host such innovations. This finding may reflect several underlying factors, including outdated infrastructure, limited budgets for upgrades, or the prioritization of traditional library services over newer, technology-focused initiatives.

This observation underscores an important challenge for public institutions: balancing the preservation of historical and cultural roles with the need to evolve in response to contemporary demands. Older libraries often serve as community anchors, rich with history and traditional resources, but they may face structural or bureaucratic hurdles that hinder their ability to incorporate new technologies. The costs of retrofitting older buildings to accommodate high-tech equipment, coupled with limited funding or resistance to change, can create barriers to modernization.

At a broader societal level, this finding highlights disparities in access to modern technology based on geographic and temporal factors. Communities served by older libraries may lack exposure to the benefits of DIHs, such as digital skills training or government support. This suggests that age-related disparities in library infrastructure could contribute to wider digital divides, limiting opportunities for communities that rely on these institutions as a primary source of technological access and education.

Understanding this dynamic has implications for policymakers and library administrators. Investments in updating or redesigning older libraries could help narrow the gap and ensure that these facilities can continue to meet the evolving needs of their communities. Additionally, this suggest a need for targeted funding and support programs that help older libraries transition to more modern service models, preserving their historical value while enabling them to serve as hubs for digital innovation.

5.4 Limitations

While this study provides clarity into the factors influencing the presence of DIHs in libraries, several limitations must be acknowledged to provide a balanced perspective.

5.4.1 Modeling Limitations

The logistic regression model used in this analysis offers a simplified representation of the complex factors influencing the establishment of DIHs. The low R^2 value indicates that a significant proportion of the variation in DIH presence remains unexplained by the variables included in the model. While this highlights the importance of the variables analyzed, it also suggests that other factors, such as community demand, administrative priorities, or external funding, were not captured.

5.4.2 Data Constraints

The dataset used in this study may have its inherent limitations. For instance, the data reflects a single snapshot in time, preventing the analysis of trends or temporal changes. Additionally, some variables, such as the year the library was built or the number of parking spaces, may serve as proxies rather than direct measures of a library's capacity to innovate. The lack of

direct data on factors like community engagement, staff expertise, or budgetary allocations limits the model's ability to explain of DIH adoption.

5.4.3 Generalizability

The study focuses on libraries within Toronto, which could limit the generalizability of the findings. For example, cultural, economic, and policy differences in other regions might lead to different relationships between the predictors and the presence of DIHs. Libraries in urban areas may face challenges distinct from those in rural areas, such as space constraints, differing user demographics, or funding opportunities, which are not explicitly addressed in this model.

5.4.4 Assumptions

The logistic regression model assumes a normal distribution of predictor variables, and linear relationships between the predictors and the presence of a DIH, which may oversimplify the real-world dynamics. Non-normal distribution or non-linear relationships were not explored, potentially omitting clarity.

5.5 Future Steps

This study provides a foundational understanding of the structural factors influencing the adoption of DIHs in libraries. However, several important questions remain unanswered, and future research should aim to narrow these gaps to enhance the understanding of the topic.

5.5.1 Incorporating Qualitative Factors

While this study focused on structural features such as parking spaces, workstations, and the age of the library, the role of qualitative factors remains unexplored. Future research should investigate how variables like community engagement, staff expertise, and administrative priorities influence DIH adoption. Surveys or interviews with library staff and patrons could uncover the motivations, challenges, and strategies behind establishing DIHs, providing richer context beyond what structural data can offer.

5.5.2 Examining External Influences

The impact of external influences, such as local government policies, regional funding availability, and partnerships with private organizations, should be investigated further. Understanding how these factors interact with the library's internal features could show additional pathways to fostering innovation.

5.5.3 Temporal Trends and Evolution

This analysis used a cross-sectional dataset, which limits the ability to capture changes over time. Longitudinal studies would allow researchers to examine how libraries adapt to technological advancements and shifting community needs.

5.5.4 Understanding User Outcomes

While this study focused on the factors that influence the presence of DIHs, future research should assess the outcomes for library users. Do DIHs improve digital literacy in the community? How do they impact job training, education, or access to technology in underserved areas? Evaluating these outcomes would help stakeholders understand the broader societal value of DIHs and justify further investment in their establishment.

5.5.5 Exploring Regional and Cultural Differences

Given that this study's findings are limited to Toronto, future research should explore cultural and regional differences in DIH adoption. International comparisons could uncover how varying economic, cultural, and policy contexts shape the adoption and success of DIHs. Similarly, libraries in rural areas might face different challenges and opportunities compared to those in urban centers.

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 follow a normal distribution, while other predictor variables might follow an exponential distribution.

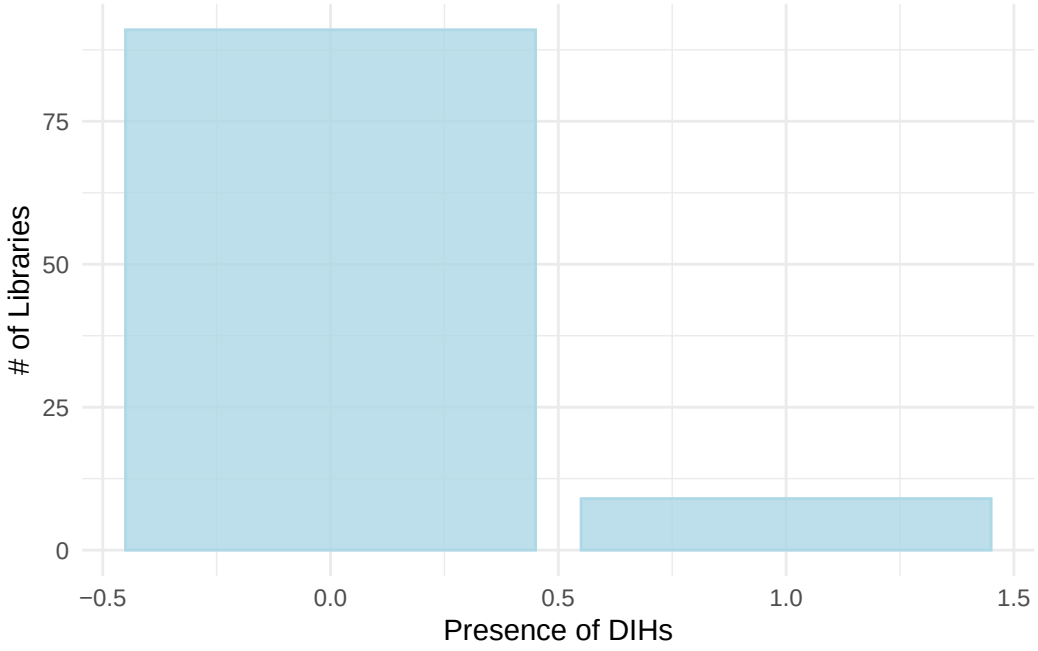

Figure 1: Presence of DIHs in Toronto Libraries

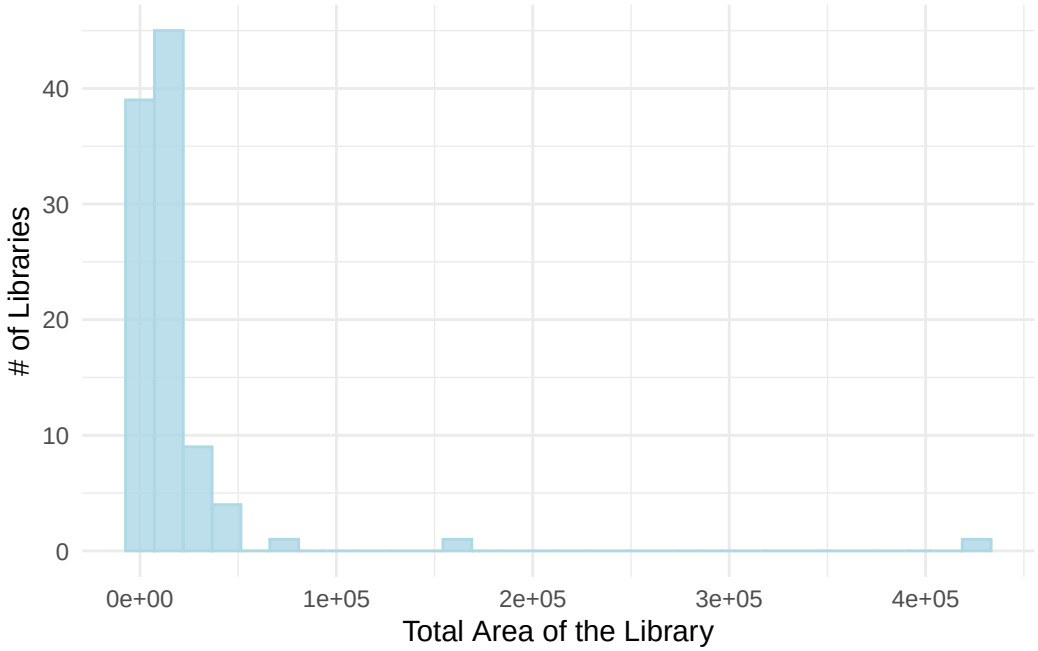

Figure 2: Distribution of Total Areas of Libraries in Toronto

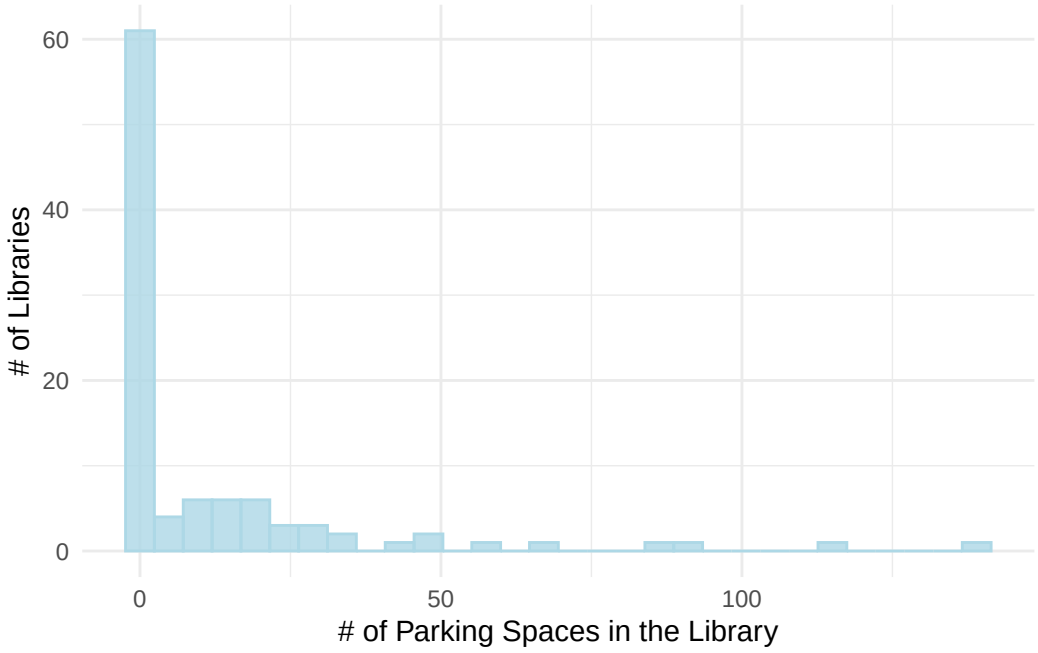

Figure 3: Distribution of Number of Parking Spaces at Libraries in Toronto

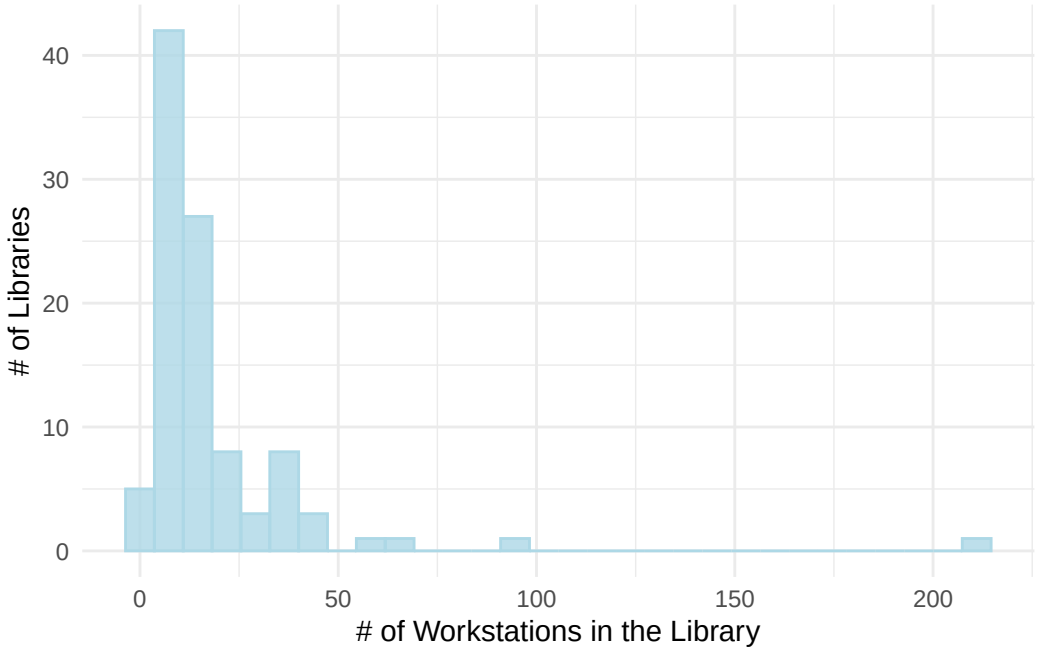

Figure 4: Distribution of Number of Workstations in Libraries in Toronto

Figure 5: Distribution of Age of Libraries in Toronto

B Model Details

B.1 Distribution of Posterior

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

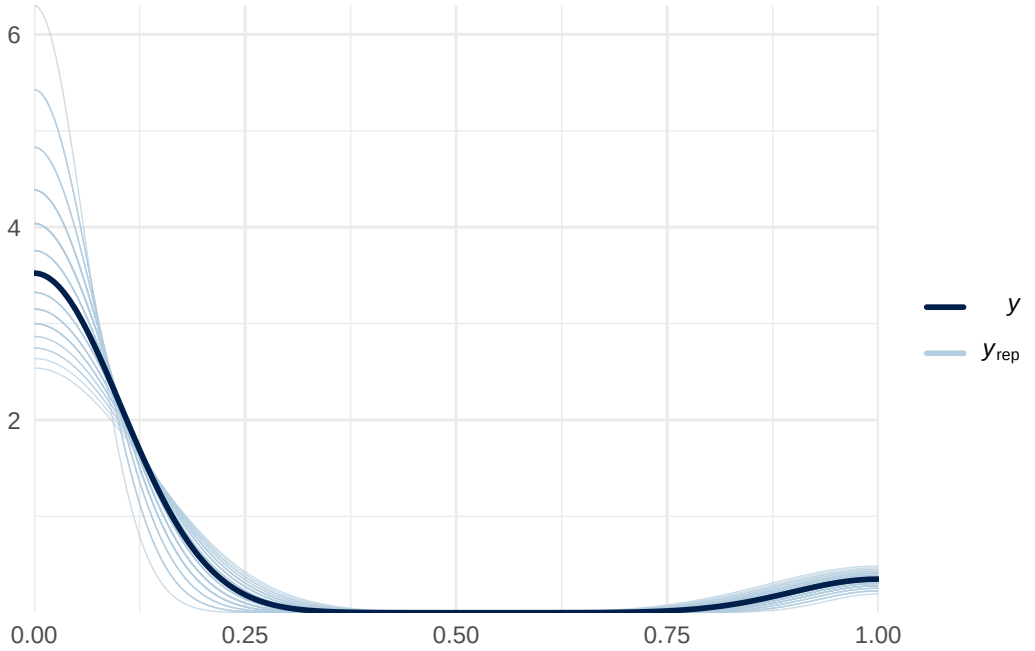

Figure 6: Posterior Distribution for the Logistic Regression Model

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 tend to have a DIH. On the opposite, the pattern shown in Figure 10 tells us older libraries tend to have no DIH.

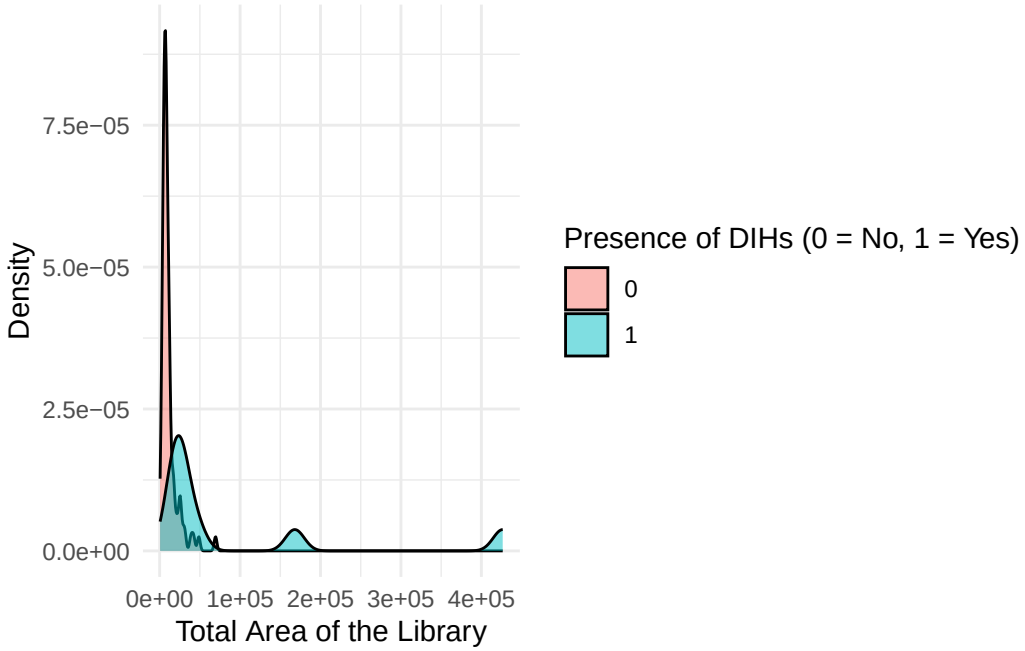

Figure 7: Density of `area` Categorized by Presence of DIHs

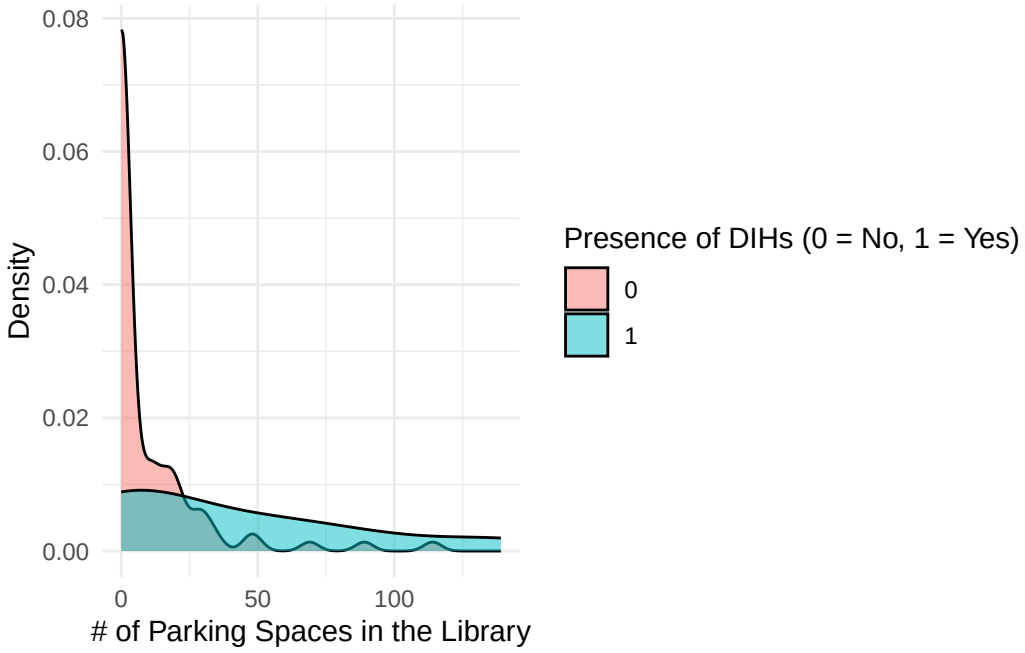

Figure 8: Density of `parking` Categorized by Presence of DIHs

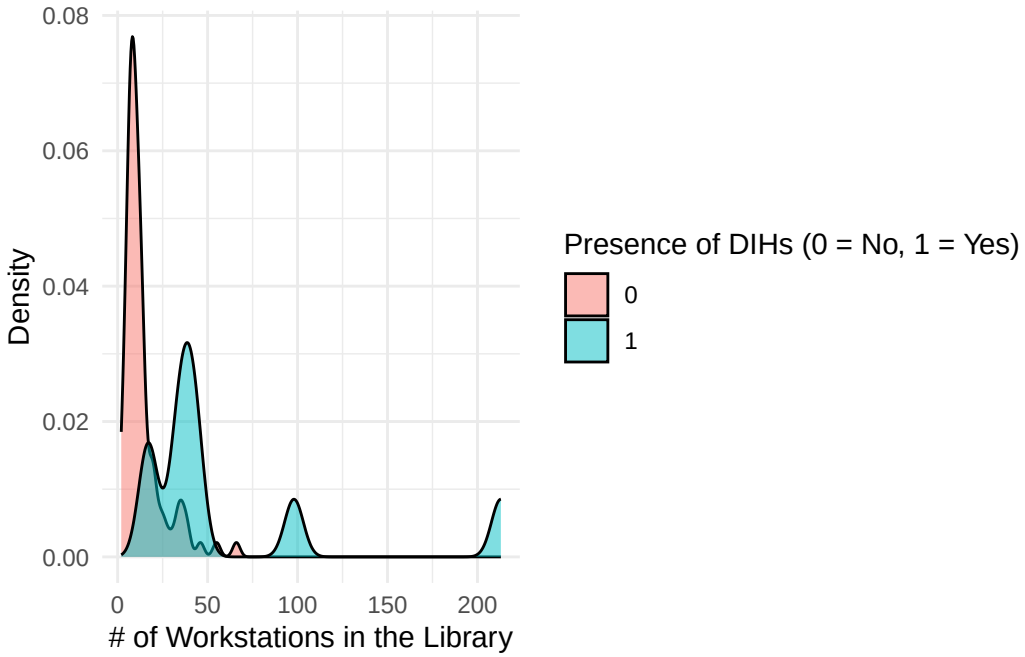

Figure 9: Density of `workstations` Categorized by Presence of DIHs

Figure 10: Density of `year` Categorized by Presence of DIHs

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.