Understanding Factors That Affects the Existence of Digital Innovation Hubs*

An Analysis of Toronto Libraries' Data

Yanzun Jiang

November 28, 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 underexplored 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. 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), `rstanarm` (Brilleman et al. 2018), `here` (Müller 2020), and `tinytable` (Arel-Bundock 2024).

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

2.1 Overview

This report uses a dataset collected by Toronto Public Library (Toronto Public Library 2024) and posted on Open Data Toronto (Gelfand 2022). Table 1 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.

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 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 $\textit{year} = 2024 - \textit{year built}$. 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.

Theeir descriptive statistics and graphs of their distribution is included in appendix (Section A).

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in **?@sec-model-details**.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \qquad (1)\\
\mu_i &= \alpha + \beta_i + \gamma_i \qquad (2)\\
\alpha &\sim \text{Normal}(0, 2.5) \qquad (3)\\
\beta &\sim \text{Normal}(0, 2.5) \qquad (4)\\
\gamma &\sim \text{Normal}(0, 2.5) \qquad (5)\\
\sigma &\sim \text{Exponential}(1) \qquad (6)
\end{aligned}
$$

We run the model in R using the `rstanarm` package of Brilleman et al. (2018). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

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.

Table 1: 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

Table 2: Descriptive Statistics of `area`

Mean	Median	Minimum	Maximum	IQR
18129.3	8496.5	554	426535	8817.8

Appendix

A Descriptive Statistics and Graphs of the Dataset

A.1 Preview of the Dataset

A.2 Descriptive Statistics

A.3 Graphs

Table 3: Descriptive Statistics of `parking`

Mean	Median	Minimum	Maximum	IQR
12	0	0	139	15.2

Table 4: Descriptive Statistics of `workstations`

Mean	Median	Minimum	Maximum	IQR
17.9	11	2	213	12

Table 5: Descriptive Statistics of `year`

Mean	Median	Minimum	Maximum	IQR
53.8	49.5	1	117	28.2

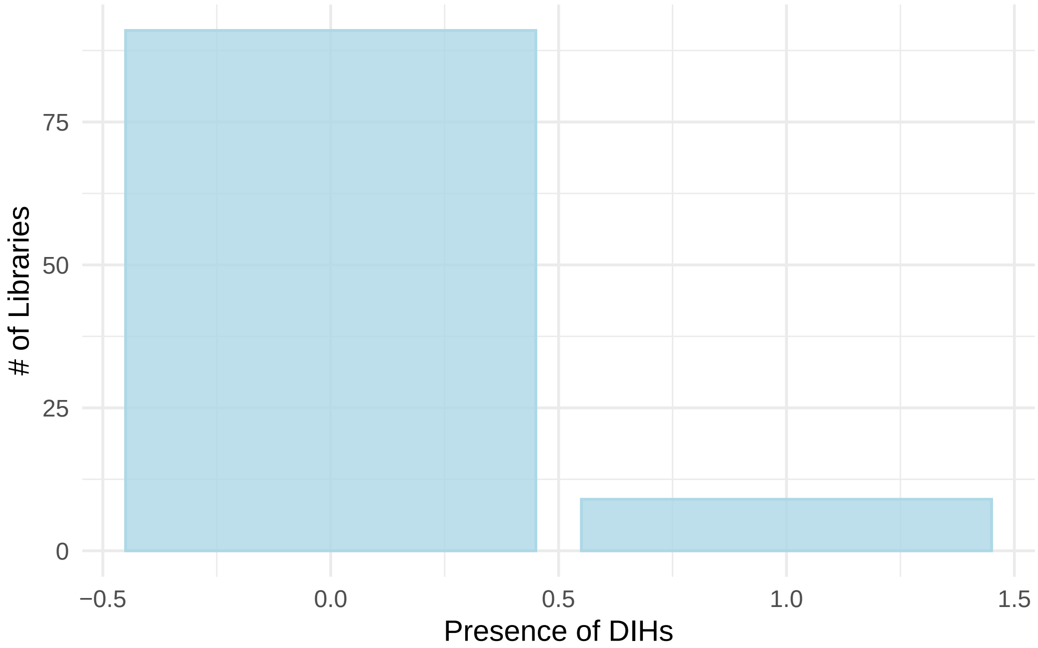

Figure 1

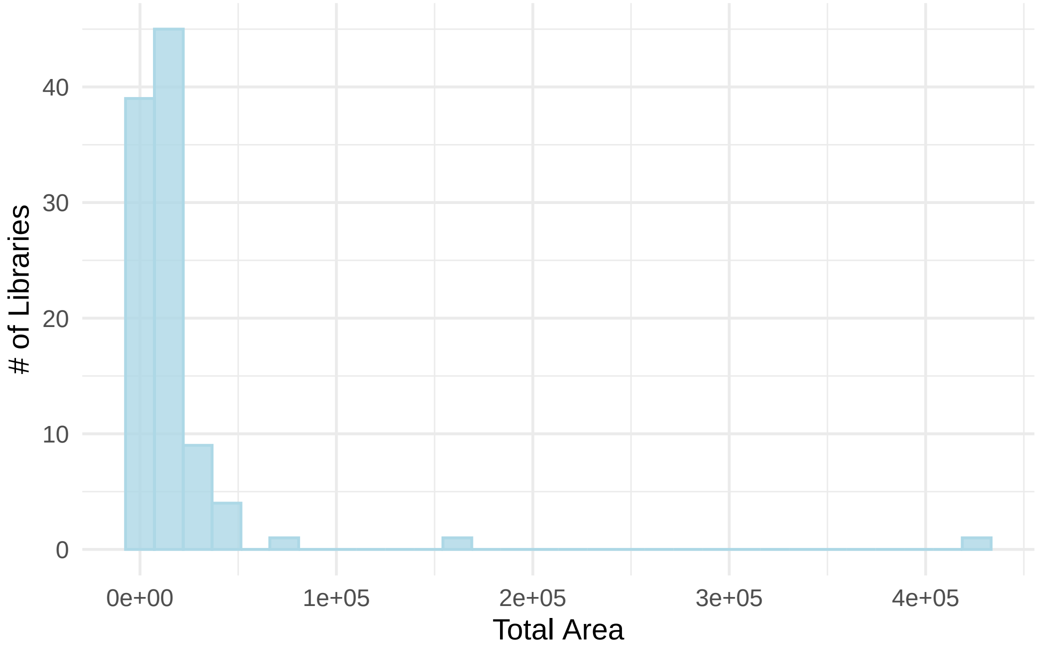

Figure 2

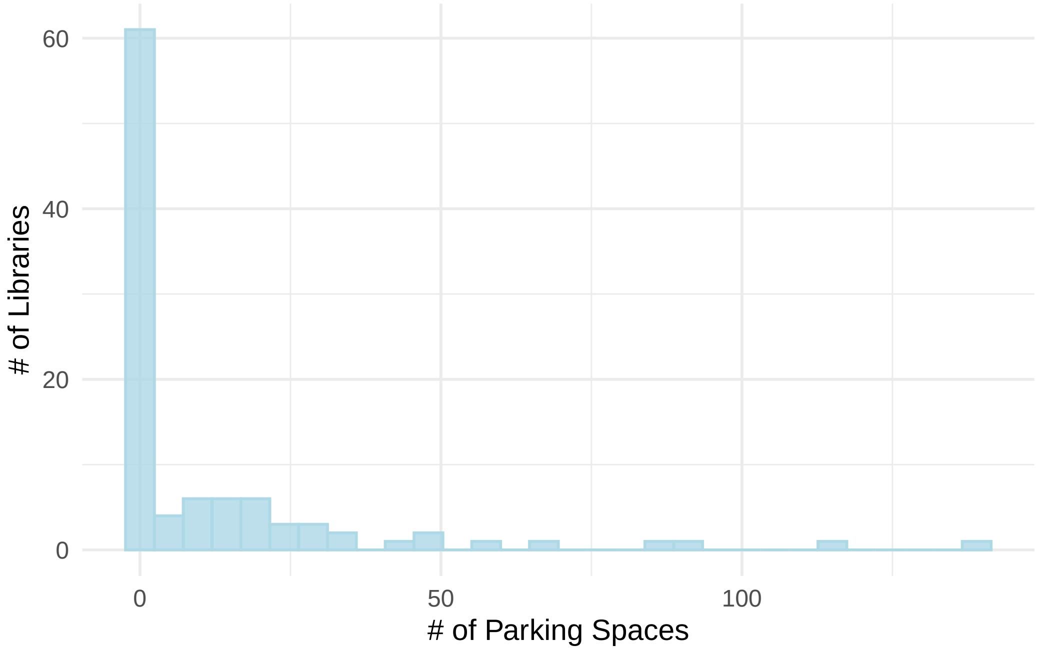

Figure 3

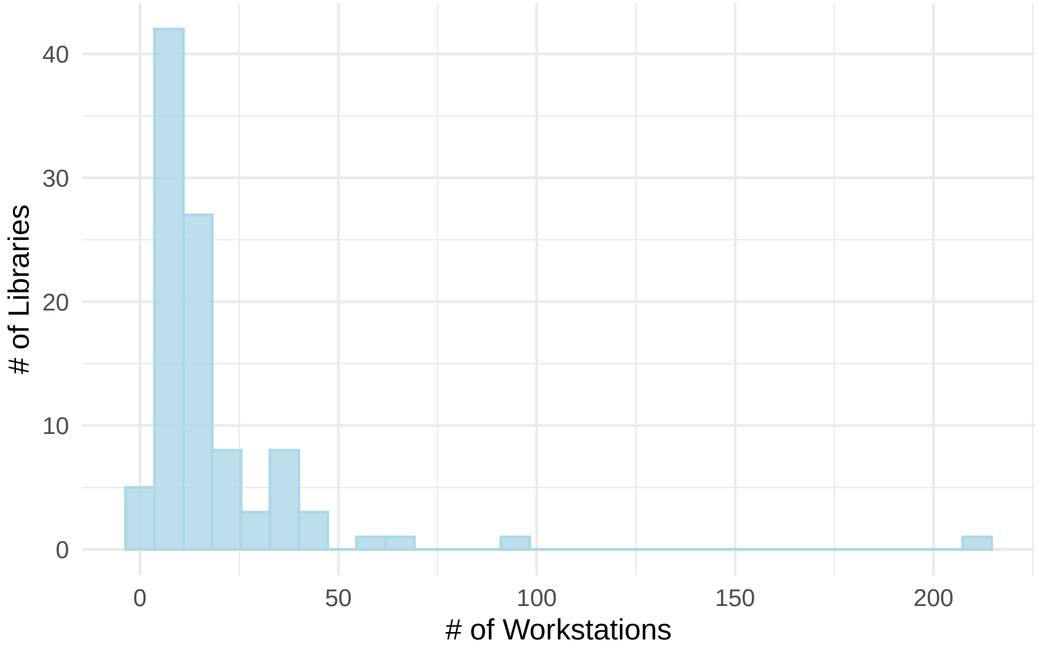

Figure 4

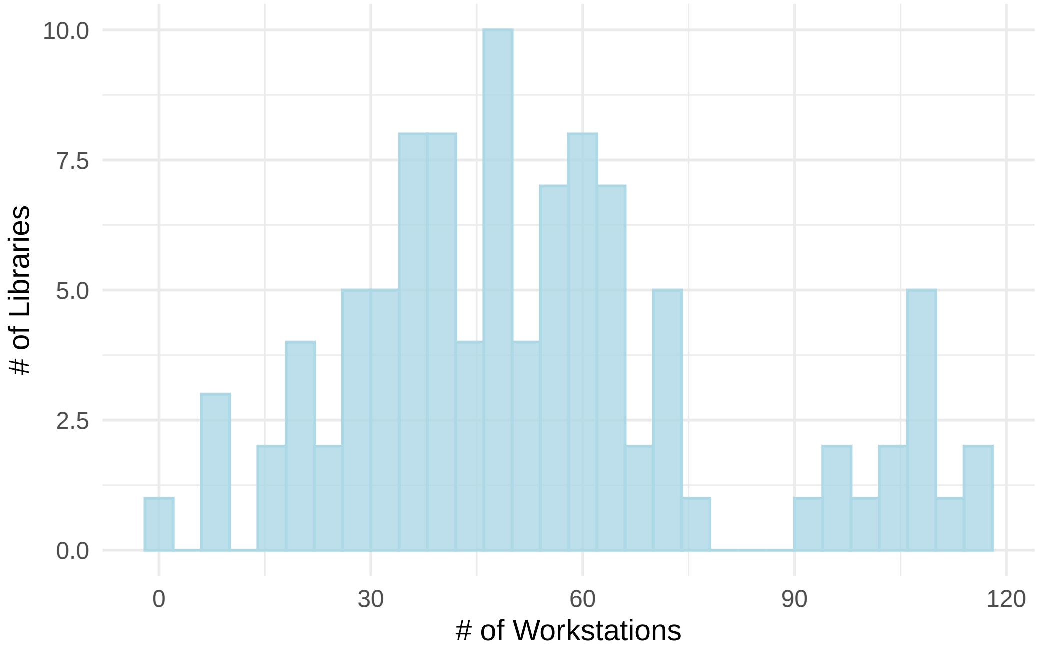

Figure 5

References

Arel-Bundock, Vincent. 2024. *Tinytable: Simple and Configurable Tables in 'HTML', 'LaTeX', 'Markdown', 'Word', 'PNG', 'PDF', and 'Typst' Formats*. https://CRAN.R-project.org/package=tinytable.

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.

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.