Ayalysis of the Ontario Public Library

Tong Su

10/20/2021

Research Question:

How does the number of visitors in the public library affected by different factors?

This research question matters for me because the library is an important study place for students. By analyzing the statistics of the public libraries, it will help us better understand the public libraries. We students could have more options to choose the library we are interested in.

Background

To fully understand my research question, I have investigated the UofT library website for papers related to determining the number of visitors to the public libraries. Unfortunately, I do not see many closely related. Although this is not a popular research area, it is still worth investigating as it is important for the people who visited the library. We could also provide more information for the latter researchers if they are also interested. From all the existing researches, I have chosen three that are related to our research question.

The first paper is about how does the expenditures and resources of public libraries relate to the teenagers and children visitors (Joo, S., & Cahill, M. 2019). This paper showed that there is a strong linear relationship between them regardless of the size of the library. The result has confirmed my research question that expenditures and resources of public libraries will affect the population served in the public library. I have also included them as the predators. What is different from this paper is that instead of focusing on children and young adults, I choose the visitors regardless of their age as the response variable.

Another related paper investigated the relationships between the content published in social media and user engagement in public libraries (Joo, S., Choi, N., Baek, T. H. 2018). This paper has shown that posts with an image and inspiring tones tend to receive more user engagement. The results make me consider that online factors may also affect the population served in the public libraries. When choosing the factors, I may also include them in my analysis.

The last paper is about the factors influencing public library visits (Huang, L.-M. and Tahamtan, I. 2018). The result has shown that the most relevant variables for predicting the number of visitors to the public libraries were total operating expenditures, the usage of public computers with the internet, audio and video physical units, and the number of children's programs. This paper is relevant to our research question because I could also include these factors in my research since they are confirmed to have a relationship with the population.

Data Source

The data set we are going to use is the Ontario public library statistics in 2020.

Website: https://data.ontario.ca/en/dataset/ontario-public-library-statistics

Table 1: Description of the main variables

Variables	Description
total_funds	Total amount of funds available in each library
number_of_resources	Total number of resources available, including physical and electronic item
number_of_visitors	Total number of visitors that have visited the library in person or online
number_of_cardholders	Total number of active library cardholders in each library
total_operating_revenues	Total amount of operating revenue generated in each library
total_operating_expenditures	Total amount of operating expenditure spent in each library
total_space_provided	Total amount of space available in each library
total_circulation_of_materials	Total number of circulations of all library materials in each library
region	Whether the library is in the northern or southern part of Ontario

The reason why I choose to include them is most of them are continuous variables. I think continuous variables are easier for us to generate linear relationships between different factors and our response variables. Furthermore, these variables do not include many zeros. If the variables include many zero values, it is hard to generate a proper linear graph since the zero values will not help determine the linear relationship. The third reason is that some of the variables included are already proven to have correlations such as the number of resources and total operating expenditures.

Lineaer model:

My research question can be answered using a linear model because the variables I use are all numerical with just one categorical variable. I will be using number of visitors as as my response variable and use the other variables as predictors because in our reaserch question, we want to predict the number of visitors in the library based on different factors.

To use linear regression for these variables, we have four assumptions in total.

1. Linearity of the Relationship
2. Covariance of the Errors
3. Common Error Variance
4. Normality of Errors

We believe the linear regression model is appropriate for these variables. For the first assumption, the variables we have chosen almost cover all characteristics of a library. The other factors will be related to these factors. In this case, we believe that only the predictors we are including are related to the response. For the second assumption, since the data for each library are collected separately, the errors should not be correlated with one another. For the third and the last assumption, we may assume that the error has a common variance and follows the normal distribution.

Exploratory data analysis

Table 2: Numerical summaries of the location and spread of the data

Variables	Min	Max	Mean	Interquantile Range	Median	Standard deviation
Total funds	0	149778587	1348282	479542	80222	8242493
Number of resources	0	20898825	214253	66047	13024	1187246
Number of visitors	0	749552	7053	1491	118	44860.05
Number of cardholders	0	813014	11465	6122	900	49704.59
Total operating revenues	248	211367610	2025768	749049	158298	11717139

Variables	Min	Max	Mean	Interquantile Range	Median	Standard deviation
Total operating expenditures	248	209434723	1975867	734957	138146	11577021
Total space provided	0	1883890	21920	12000	2575	107947.7
Total circulation of materials	0	420019	6021	1648	211	30209.9

From the location of the data, we could observe that most of the data have minimum 0 and large maximum values. This will affect the mean as well as the spread of the data. We could also observe that there is a huge gap between the mean and median values of the data. This is because the mean is not robust to outliers but the median is more robust. From the spread of the data, we could observe that the range, interquartile range, and standard deviation for each numerical variable are quite large. This is because our data contain some extreme values which affect the spread of our variables.

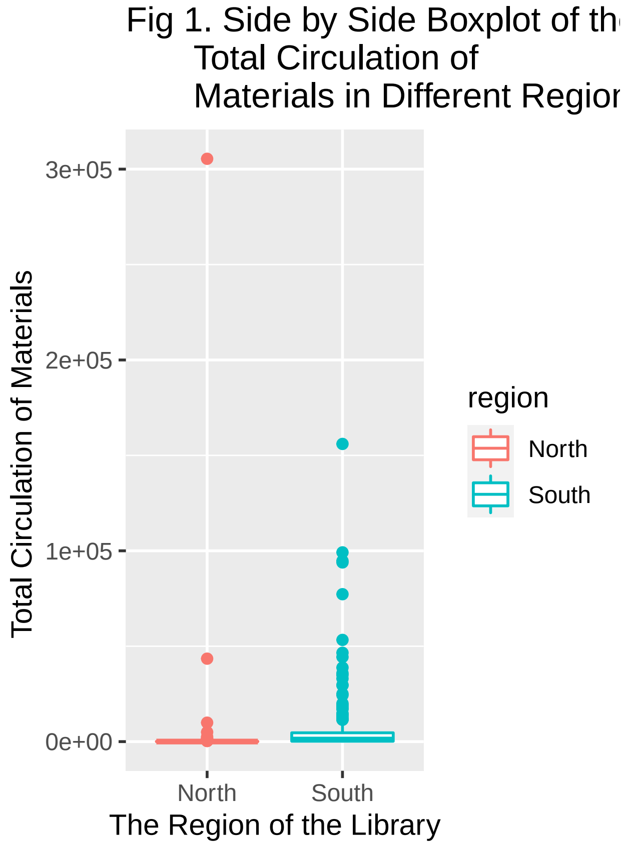
Fig 1. Side by Side Boxplot of th
Total Circulation of
Materials in Different Regior

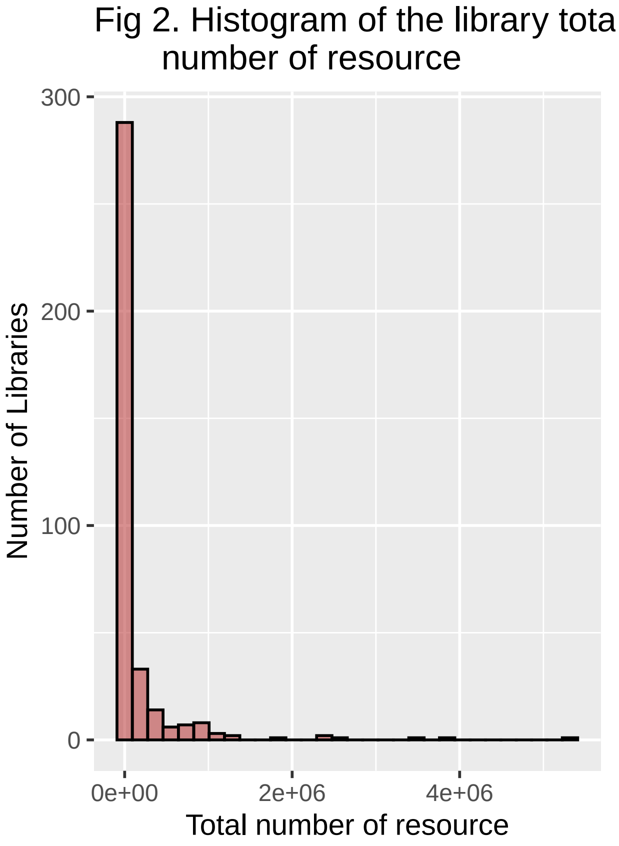
Fig 2. Histogram of the library tota
number of resource

Fig 1. is the side-by-side boxplot of the total circulation of materials in a different region. From the graph, we observe that for both regions, the boxplots are extremely right-skewed. This is because there are some extreme outliers with an extremely high total circulation of materials. We could observe from the skewed graph that the median total circulation of materials is higher in the southern area than in the north. The total circulation of materials also has a larger interquartile range than the north. However, The north total circulation of materials has a wider range than the south.

Fig 2. is a histogram of the library's total number of resources. From the graph, we could observe that this graph is extremely right-skewed. This indicates that the mean value of the total number of resources is greater than its median. This result follows the data we obtained in our numerical summaries. The graph is unimodal with only one mode around 0. The range of the total number of resources is very large from 0 to over 40000000. Due to the presence of extreme values, the mean value of the expenditure on consumption is greatly impacted since it is not robust.

Reference

1. Joo, S., & Cahill, M. (2019). The relationships between the expenditures and resources of public libraries and children's and young adults' use: An exploratory analysis of Institute of Museum and Library Services public library statistics data. Journal of Librarianship and Information Science, 51(2), 307–316. https://doi.org/10.1177/0961000617709057

2. Joo, S., Choi, N., Baek, T. H. (2018). Library marketing via social media: The relationships between Facebook content and user engagement in public libraries, 42(6), 940–955. https://doi.org/10.1108/OIR-10-2017-0288

3. Huang, L.-M. and Tahamtan, I. (2018). Why do People Come? The Factors Influencing Public Library Visits. In L. Freund (Ed.), Proceedings of the Association for Information Science and Technology (pp. 832– 833.) Hoboken, NJ: Wiley. https://doi.org/10.1002/pra2.2018.14505501136

4. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

5. Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://CRAN.R-project.org/package=janitor

6. Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2021). skimr: Compact and Flexible Summaries of Data. R package version 2.1.3. https://CRAN.R-project.org/package=skimr

7. Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.

8. Ontario Data Catalogue. (2021). Ontario public library statistics, 2020 Retrieved from https://data.ontario.ca/en/dataset/ontario-public-library-statistics/resource/c897cb92-d90f-4965-b2d1-c5b08569bb58

9. Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot