Analysis of the trends that lead to the survival & downfall of American public libraries

Elizabeth Edminster & Sarah Evans

2024-03-28

Project Description This project aims to critically analyze data in the domain of public policy & business strategies for NGOs, in this case, public libraries. Our guiding questions are as follows: What quantifiable library traits correlate well with success over multiple years? What traits & features often lead to the demise of a library? How do library traits differ by geographical region & city population? How have libraries as a whole changed over the years from 2006-2021?

Data Description and Links Data: https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey

This data is from the Institute of Museum & Library Services. They have data on these libraries nationally from 2006 to 2021. The selected columns for our preliminary exploration are listed below. Further columns & information about the data can be found at the provided link.

The following columns are the variables selected for this project:

(NOTE) The variables considered for analysis in this data have changed slightly from 2006-2021.The following variables are the variables taken from 2021. If a given variable does not exist in prior years, it will not be considered, but no extra previously removed columns will be added.

LIBID - Identifier: Library identification code assigned by the state. IMLS assigns the FSCSKEY to this field if the state did not assign a code./

LIBNAME - Identifier: Name of library (administrative entity)/

STABR - Identifier: Two-letter American National Standards Institute (ANSI) State Code. (See Appendix D for list of State Codes.)/

C_LEGBAS - Categorical: Legal Basis Code CC–City/County CI–Municipal Government (city, town, or village) CO–County/Parish LD–Library District MJ–Multi-jurisdictional NL–Native American Tribal Government NP–Non-profit Association or Agency SD–School District OT–Other/

POPU_LSA - Numeric: Population of the Legal Service Area -1–Missing -3–Temporarily closed administrative entity -9–Data suppressed for analytic purposes/

VISITS - Numeric: This is the total number of persons entering the library for whatever purpose during the year./

BKVOL - Numeric: Books in print. Books are non-serial printed publications (including music scores or other bound forms of printed music, and maps) that are bound in hard or soft covers, or in loose-leaf format. Does not include unbound sheet music. Includes non-serial government documents. Including duplicates./

EBOOK - Numeric: E-books are digital documents (including those digitized by the library), licensed or not, where searchable text is prevalent, and which can be seen in analogy to a printed book (monograph). E-books are loaned to users on portable devices (e-book readers) or by transmitting the contents to the user's

personal computer for a limited time. Includes ebooks held locally and remote e-books for which permanent or temporary access rights have been acquired. Including duplicates./

CAPITAL - Numeric: Report major capital expenditures (the acquisition of or additions to fixed assets). Examples include expenditures for (a) site acquisitions; (b) new buildings; (c) additions to or renovation of library buildings; (d) furnishings, equipment, and initial book stock for new buildings, building additions, or building renovations; (e) library automation systems; (f) new vehicles; and (g) other one-time major projects. Includes federal, state, local, or other revenue used for major capital expenditures. Only funds that are supported by expenditure documents (e.g., invoices, contracts, payroll records, etc.) at the point of disbursement should be included. Estimated costs are not included. Excludes expenditures for replacement and repair of existing furnishings and equipment, regular purchase of library materials, and investments for capital appreciation./

PRMATEXP - Numeric: Report all operating expenditures for the following print materials: books, current serial subscriptions, government documents, and any other print acquisitions.

ELMATEXP - Numeric: Report all operating expenditures for electronic (digital) materials. Types of electronic materials include e-books, audio and video downloadables, e-serials (including journals), government documents, databases (including locally mounted, full text or not), electronic files, reference tools, scores, maps, or pictures in electronic or digital format, including materials digitized by the library/

STAFFEXP - Numeric: This is the sum of Salaries & Wages Expenditures and Employee Benefits Expenditures/

STATNAME - Categorical: Name Change Code 00–No change from last year 06–Official name change 14–Minor name change/

LOCALE_MOD - Categorical: Urban-centric locale code. The geographic location in terms of the size of the community in which it is located and the proximity of that community to urban and metropolitan areas. Assigned based on the modal locale code of associated stationary outlets (i.e., central and branch libraries). 11–City, Large: Territory inside an urbanized area and inside a principal city with population of 250,000 or more. 12–City, Mid-size: Territory inside an urbanized area and inside a principal city with a population less than 250,000 and greater than or equal to 100,000. 13–City, Small: Territory inside an urbanized area and inside a principal city with a population less than 100,000. 21–Suburb, Large: Territory outside a principal city and inside an urbanized area with population of 250,000 or more. 22–Suburb, Mid-size: Territory outside a principal city and inside an urbanized area with a population less than 250,000 and greater than or equal to 100,000. 23–Suburb, Small: Territory outside a principal city and inside an urbanized area with a population less than 100,000. 31–Town, Fringe: Territory inside an urban cluster that is less than or equal to 10 miles from an urbanized area. 32–Town, Distant: Territory inside an urban cluster that is more than 10 miles and less than or equal to 35 miles from an urbanized area. 33–Town, Remote: Territory inside an urban cluster that is more than 35 miles from an urbanized area. 41–Rural, Fringe: Census-defined rural territory that is less than or equal to 5 miles from an urbanized area, as well as rural territory that is less than or equal to 2.5 miles from an urban cluster. 42–Rural, Distant: Census-defined rural territory that is more than 5 miles but less than or equal to 25 miles from an urbanized area, as well as rural territory that is more than 2.5 miles but less than or equal to 10 miles from an urban cluster./

WEBVISIT - Numeric: Total visits (sessions) to library website -1–Missing -3–Temporarily closed administrative entity -4–Not applicable/

WIFISESS - Numeric: Total annual wireless sessions provided by the library wireless service -1–Missing -3–Temporarily closed administrative entity/

PITUSR - Numeric: Uses of public Internet computers per year -1–Missing -3–Temporarily closed administrative entity/

LOANTO - Numeric: Total annual loans provided to other libraries -1–Missing -3–Temporarily closed administrative entity/

LOANFM - Numeric: Total annual loans received from other libraries -1–Missing -3–Temporarily closed administrative entity/

TOTCIR - Numeric: Total annual circulation transactions -1–Missing -3–Temporarily closed administrative entity/

ELMATCIR - Numeric: Use of Electronic Materials – The total annual circulation of all electronic materials -1–Missing -3–Temporarily closed administrative entity

PHYSCIR - Numeric: Physical item circulation – The total annual circulation of all physical library materials of all types, including renewals. -1–Missing -3–Temporarily closed administrative entity

REGBOR - Numeric: Registered Users -1–Missing -3–Temporarily closed administrative entity

Data Exploration The total size of this data is 9215 observations. We will be looking at 23 variables.

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'dplyr' was built under R version 4.3.2

## Warning: package 'stringr' was built under R version 4.3.2

## Warning: package 'lubridate' was built under R version 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data <- read.csv("C:/Users/lized/OneDrive/Documents/GitHub/R-Analysis-Project/pls_fy2021_csv/PLS_FY2021

#This data specifically is the 2021 data
df <- data %>% select(c(LIBID, LIBNAME, STABR, C_LEGBAS, POPU_LSA, VISITS, BKVOL, EBOOK, CAPITAL, PRMAT
df$LOCALE_MOD <- as.factor(df$LOCALE_MOD)
df$C_LEGBAS <- as.factor(df$C_LEGBAS)
df$STATNAME <- as.factor(df$STATNAME)
summary(df)
```

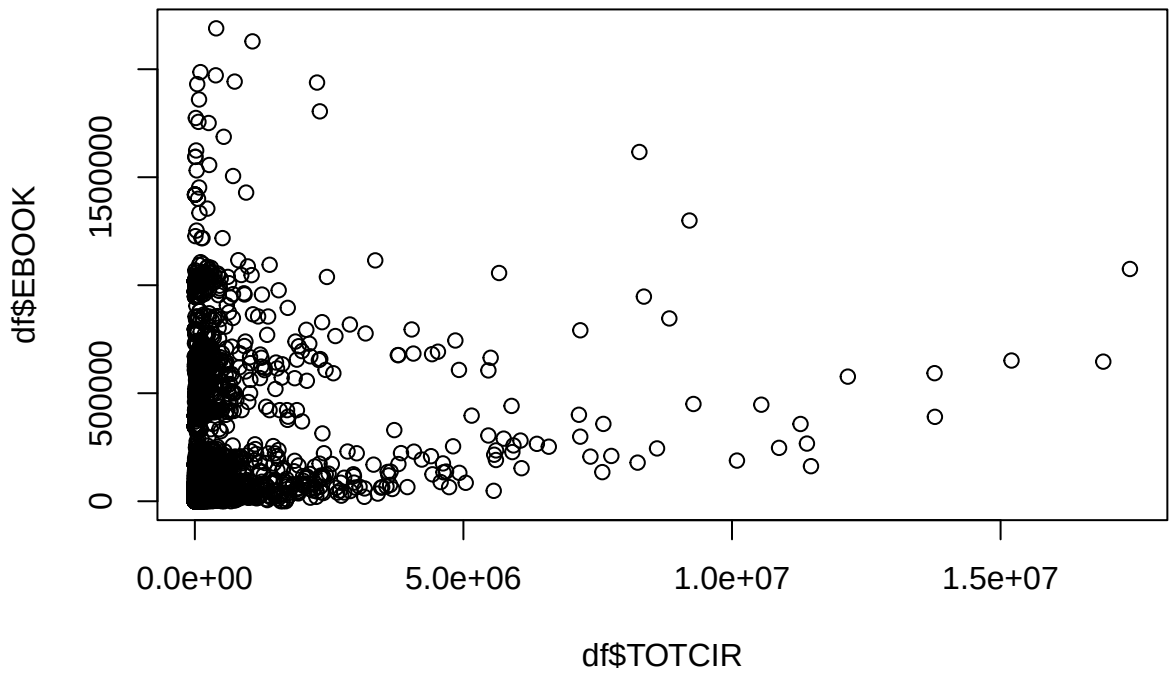
```
##     LIBID             LIBNAME             STABR             C_LEGBAS
##  Length:9215        Length:9215        Length:9215        CI     :4850
##  Class :character   Class :character   Class :character   LD     :1405
##  Mode  :character   Mode  :character   Mode  :character   NP     :1328
##                                                           CO     : 923
##                                                           MJ     : 286
##                                                           SD     : 180
##                                                           (Other): 243
##     POPU_LSA           VISITS            BKVOL              EBOOK
```

```
##  Min.   :     -9   Min.   :     -3   Min.   :      -3   Min.   :     -3
##  1st Qu.:   2198   1st Qu.:   3250   1st Qu.:   13698   1st Qu.:  14555
##  Median :   7280   Median :  11000   Median :   26873   Median :  41154
##  Mean   :  35657   Mean   :  45306   Mean   :   71831   Mean   : 113585
##  3rd Qu.:  22831   3rd Qu.:  36216   3rd Qu.:   60241   3rd Qu.: 105022
##  Max.   :4507419   Max.   :6722578   Max.   :22168629   Max.   :2189199
##
##     CAPITAL            PRMATEXP           ELMATEXP           STAFFEXP
##  Min.   :      -3   Min.   :      -3   Min.   :      -3   Min.   :       -9
##  1st Qu.:       0   1st Qu.:    6170   1st Qu.:     375   1st Qu.:       -9
##  Median :       0   Median :   16663   Median :    2653   Median :   156343
##  Mean   :  157210   Mean   :   72254   Mean   :   64566   Mean   :   954892
##  3rd Qu.:   12260   3rd Qu.:   51356   3rd Qu.:   19162   3rd Qu.:   578944
##  Max.   :73338263   Max.   :10764492   Max.   :13555446   Max.   :211582281
##
##  STATNAME    LOCALE_MOD      WEBVISIT           WIFISESS
##  0 :9140   42     :1957   Min.   :      -4   Min.   :      -3
##  6 :  20   21     :1877   1st Qu.:       0   1st Qu.:       0
##  14:  55   43     :1580   Median :    5508   Median :    1560
##            32     :1122   Mean   :  123717   Mean   :   25142
##            33     : 677   3rd Qu.:   36909   3rd Qu.:    8245
##            41     : 508   Max.   :22722256   Max.   :27836110
##            (Other):1494
##      PITUSR            LOANTO             LOANFM           TOTCIR
##  Min.   :     -3   Min.   :    -3.0   Min.   :    -3   Min.   :      -3
##  1st Qu.:    306   1st Qu.:    31.5   1st Qu.:    50   1st Qu.:    7546
##  Median :   1090   Median :   548.0   Median :   593   Median :   25986
##  Mean   :   5751   Mean   :  6969.2   Mean   :  7153   Mean   :  168778
##  3rd Qu.:   3610   3rd Qu.:  5436.5   3rd Qu.:  4832   3rd Qu.:   90500
##  Max.   :1484987   Max.   :706516.0   Max.   :919237   Max.   :17408320
##
##     ELMATCIR           PHYSCIR            REGBOR
##  Min.   :      -3   Min.   :      -3   Min.   :     -3
##  1st Qu.:     978   1st Qu.:    5880   1st Qu.:    990
##  Median :    4652   Median :   20213   Median :   3136
##  Mean   :   50362   Mean   :  118416   Mean   :  17293
##  3rd Qu.:   20172   3rd Qu.:   68300   3rd Qu.:  10131
##  Max.   :12223192   Max.   :12717585   Max.   :2696713
##
```

```
head(df[,1:5], 5)
```

```
##        LIBID                        LIBNAME STABR C_LEGBAS POPU_LSA
## 1 AK0001-002  ANCHOR POINT PUBLIC LIBRARY    AK       NP     2123
## 2 AK0002-011     ANCHORAGE PUBLIC LIBRARY    AK       CO   288970
## 3 AK0003-002   ANDERSON COMMUNITY LIBRARY    AK       CI      275
## 4 AK0006-002 KUSKOKWIM CONSORTIUM LIBRARY    AK       MJ     6179
## 5 AK0007-002      BIG LAKE PUBLIC LIBRARY    AK       CO     6942
```

```
plot <- plot(df$TOTCIR, df$EBOOK)
```




```
plot
```

```
## NULL
```

Analysis Plan We plan to use ANOVA to compare the mean number of EBOOKS, the mean number of BKVOL, and the mean number of TOTCIR at libraries from different regions. We plan to use Linear regression to predict the capital based on the given numeric qualities. We want to see which qualities result in the most accurate model without overfitting. We plan to use classification modeling to predict whether a library would have $>$ or $<$ the mean number of total circulation. We will use RMSE to evaluate quality