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

Elizabeth Edminster & Sara Evans

2024-05-01

1.0 Introduction

1.1 Project Description This project aims to critically analyze data within the domain of public policy and business strategies, with a focus on NGOs, particularly public libraries.

We aim to address various questions through our analysis such as, what quantifiable library traits are correlated with a library's long-term success and which traits commonly contribute to the demise of a library? We consider it crucial to identify quantifiable traits associated with library success as strategies can be developed based on these factors to help libraries to better serve their communities, while also acknowledging the factors that impact a library's declining state. Moreover, we aim to explore whether library characteristics vary based on geographical location and city population, along with examining the evolution of libraries from 2011 to 2021. It's necessary to address these questions understand how libraries are distributed across different regions and whether all regions have equal access to a library. Additionally, tracking how libraries have evolved over the course of a decade will provide insights on how their community's needs have changed, as well as the popularity of libraries in general.

A library's success will be based on a combination of factors such as total inventory books, including physical books, e-books, and audio books, as this demonstrates a library's ability to provide resources to their community. In addition, the total number of visits a library receives each year will also determine success as a large number of visits indicates popularity among it's audience.

1.2 Background This data, the Public Library Survey, is run by the Institute of Museum & Library services. The data sets are publicly accessible & provide operational data on almost 9000 libraries across the US & it's territories. Data is collected annually, and the most recent data set is from 2021.

The data collected encompasses many aspects of library operation. Some examples are geographical data, such as the population of the library's service area, collection data, such as the quantity of EBooks, and administrative data, such as information on worker's salaries.

This data is collected voluntarily from libraries in a census style. In 2021, 8,903 of the 9,207 libraries responded. NY had the highest quantity of respondents at 756 responding libraries. 6 areas, such as Hawaii & Guam, had only 1 respondent.

When referring to the population that a library represents, we call it a Legal Service Area. When referring to the type of population (city, county, school, etc.), we call it a Legal Basis Code.

2.0 Data Description

Data: https://www.imls.gov/research-evaluation/data-collection/publiclibraries-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 1 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/

Modifications made to the data includes the removal of a significant amount of outliers, as well as NA and values that represented missing information (-1), temporarily close administrative entity (-3), data suppressed for analytic purposes (-9), and not applicable (-4).

3.0 Analysis

3.1 Clustering This statistical technique included K-Means clustering and model-based clustering. For the K-Means and model-based clusterings, the following variables were used: OBEREG, LIBRARIA, BKVOL, EBOOK, Total_Audio_Books, Total_Videos, REGBOR, and VISITS. The variable Total_Audio_Books was calculated using the AUDIO_PH and AUDIO_DL and the variable Total_Video was calculated using the VIDEOS_PH and AUDIO_DL.

Two datasets were used, one from 2011 and another from 2021 with a subset being taken from each to select the variables being used for the clustering analysis. The 2011 dataset orginally had 9,315 rows while the 2021 dataset had 9,215 rows. After removing a large number of outliers and omitting na values, the 2011 dataset had 6056 rows and the 2021 dataset contained 5,783 rows. Following the removal of outliers, the all of the variable in both dataset were scaled expect for OBEREG which denotes the region.

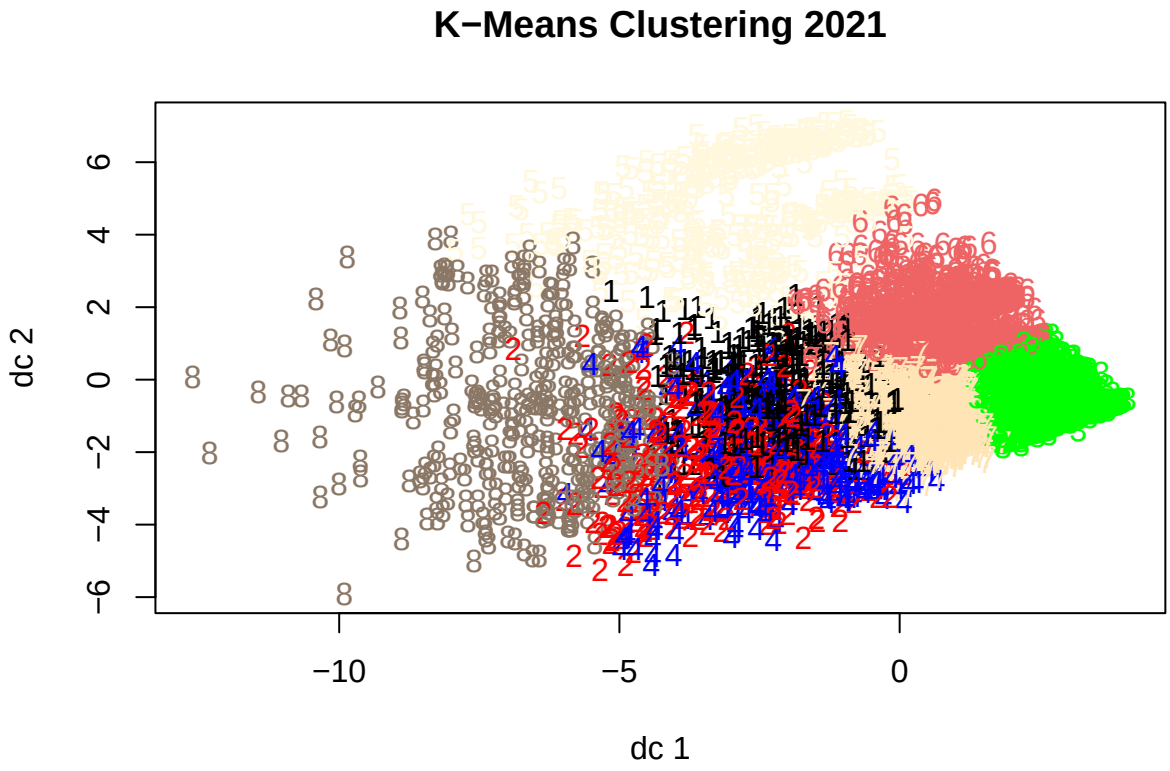
3.1.1 Clustering Results

K-Means:

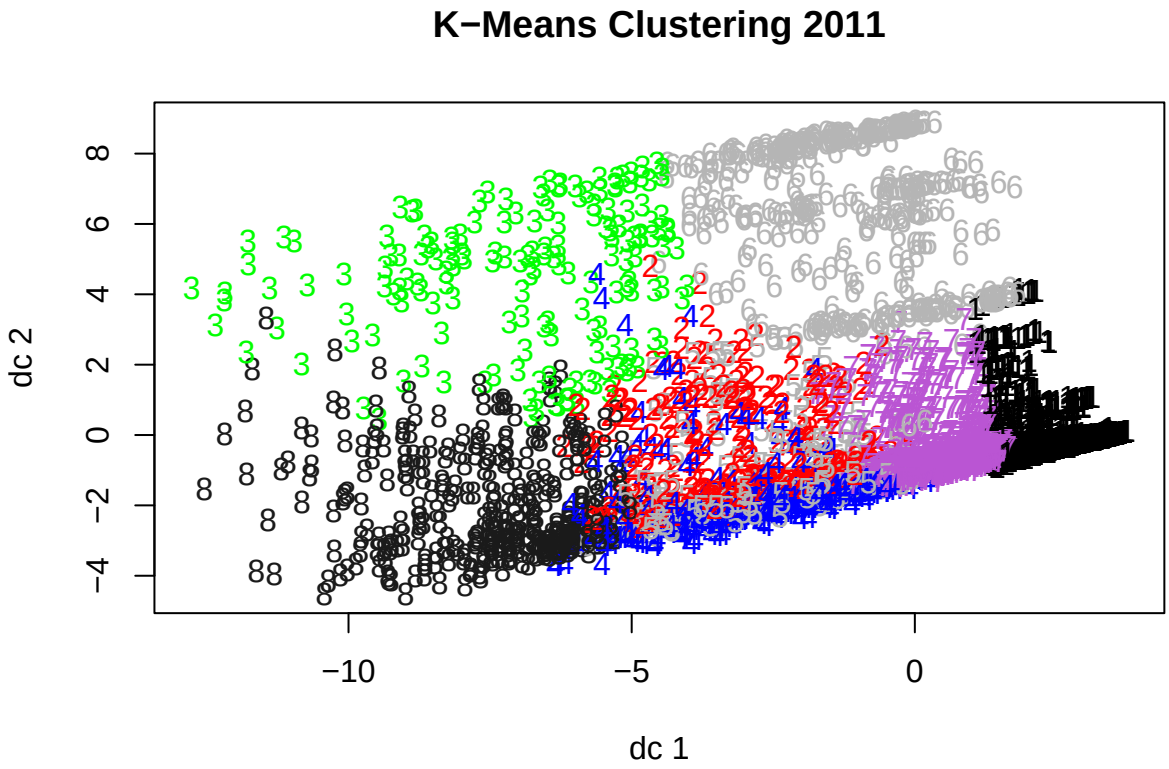
```
fit.km5 <- kmeans(cleaned_data_21.scaled, 8, nstart=25)
fit.km4 <- kmeans(cleaned_data_11.scaled, 8, nstart=25)
```

```
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
```

```
plotcluster(cleaned_data_21.scaled, fit.km5$cluster, main = "K-Means Clustering 2021")
```




```
plotcluster(cleaned_data_11.scaled, fit.km4$cluster, main = "K-Means Clustering 2011")
```

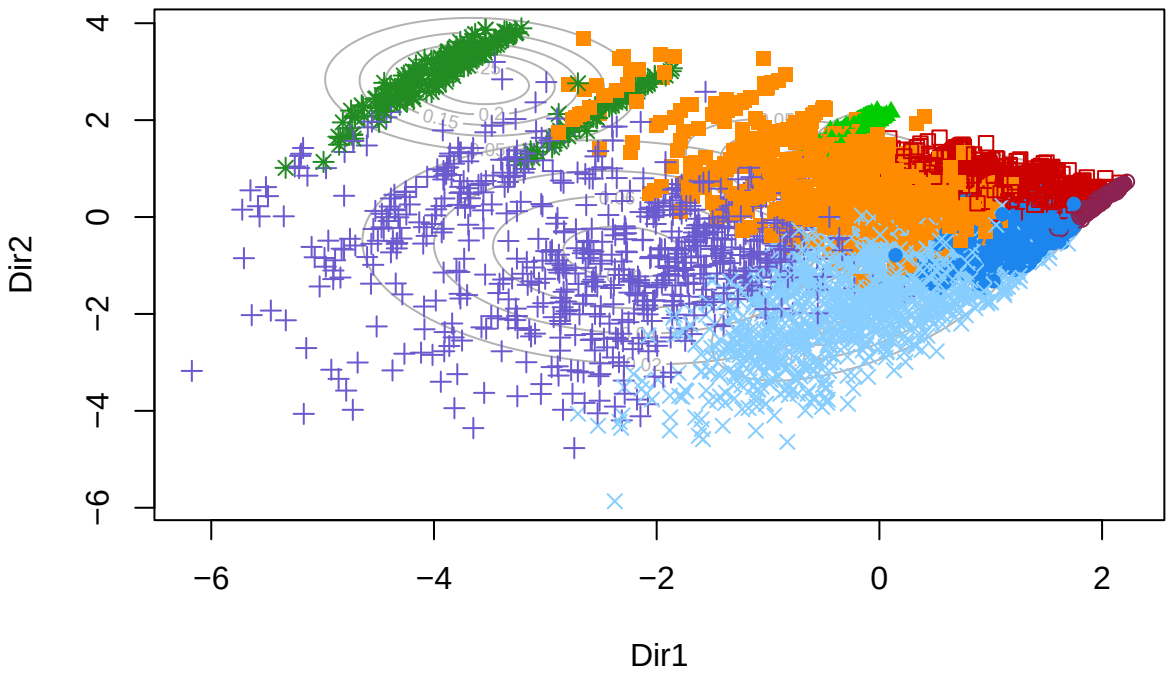


For the results of the 2021 clustering, there appears to be fairly distinct clusters, most with no other points from other clusters overlapping. Additionally, the clusters are relatively compact therefore this suggests minimal disbursement. For the most part, the clusters have close to the same size of data points indicating the clusters are balanced. Overall, it seems that the variables include in the most are relevant for distinguishing between different groups of the data.

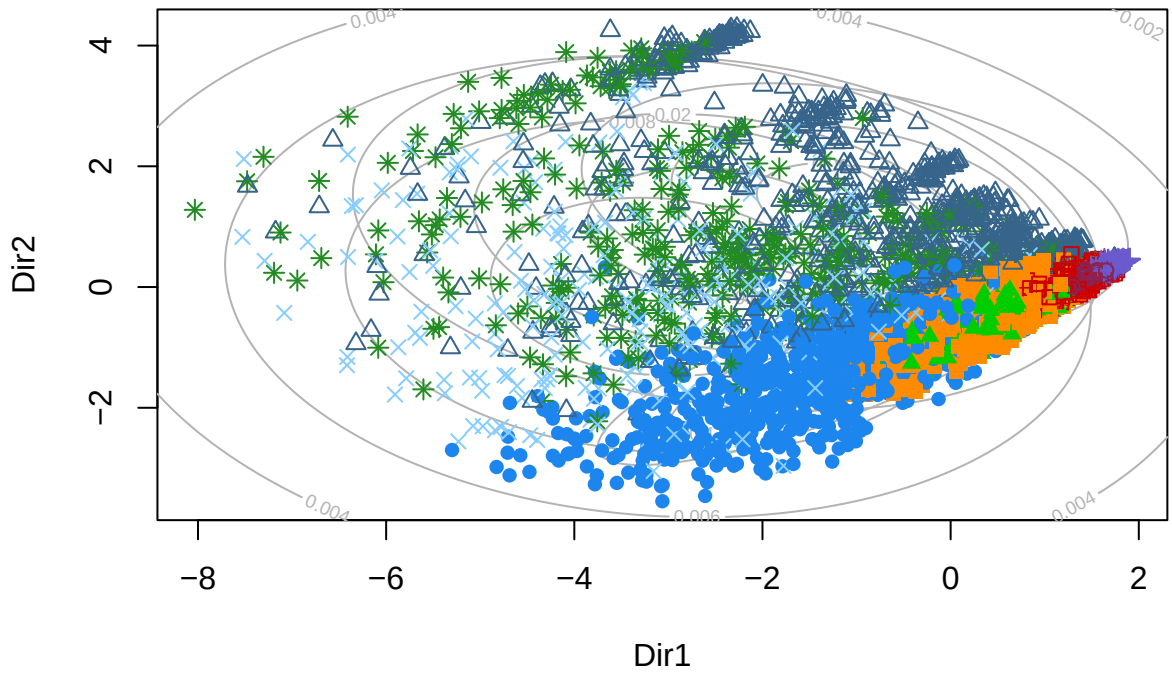
For the results of the 2011 clustering, only four clusters were used to sort the data with minimal overlapping into other clusters. However, it appears this model it much less compact than the previous model which suggests changes in business dynamics over 10 years. These cluster's aren't as balanced as the 2021 model, but they aren't too off-balanced. Overall, this model suggests that many of the variable are relevant for distinguishing groups of the data, however over the course of 10 years not all of the variables relevant in the 2021 model may be as relevant in the 2011 model.

Model-based:

```
model.dr_21 <- MclustDR(model.clus_21)
plot(model.dr_21, what = "contour")
```




```
model.dr_11 <- MclustDR(model.clus_11)
plot(model.dr_11, what = "contour")
```



For the 2021 model, there are 9 different clusters, although it's difficult to distinguish some of them from others as many appear to overlap. For the most part the clusters appear to be relatively compact meaning there's not too much disbursement. However, the clusters a highly unbalanced with two or three clusters having most of the data and the other clusters having much smaller amounts. Overall, it appears that it's difficult to identify differences among libraries in different regions as the region a library is in does not seem to correlate to the amount of visits they receive or amount of books they have.

For the 2011 model, there are also 9 different clusters that appear to be even more spread out than the 2021 model. Some of these clusters are compact while others are not, although it appears the clusters with larger amounts of data are more spread out. Additionally, these clusters also appear to be unbalanced with two to three clusters having most of the data. Similarly to the 2021 model, it libraries don't appear to be distinguishable based on the region they reside in making it difficult to judge any differences in variables over the years.

3.2 We used linear regression to identify traits that would significantly impact one of 3 dependent variables: Number of Visits (VISITS), The total circulation of books (TOTCIR), and The number of registered users (REGBOR)

First, we began by fitting a model for each of the adjusted dependent variables: Visits, Circulation, & Users using every selected trait. The adjustment was dividing the number of visits/circulation/users by the number of people in the library's given legal service area.This was done because we assume that the quantity of visits/circulation/users is proportional to the quantity of people the library represents. Then, for each of the models, we made a second model that included only the significant variables. This is based on the most recent data: 2021, and the data from a decade prior: 2011.

We removed rows where the population, the number of visits, the circulation, or the number of users was missing. This removes 45 rows in 2021, bringing our data from 9215 to 9170.This removes 55 rows in 2011,

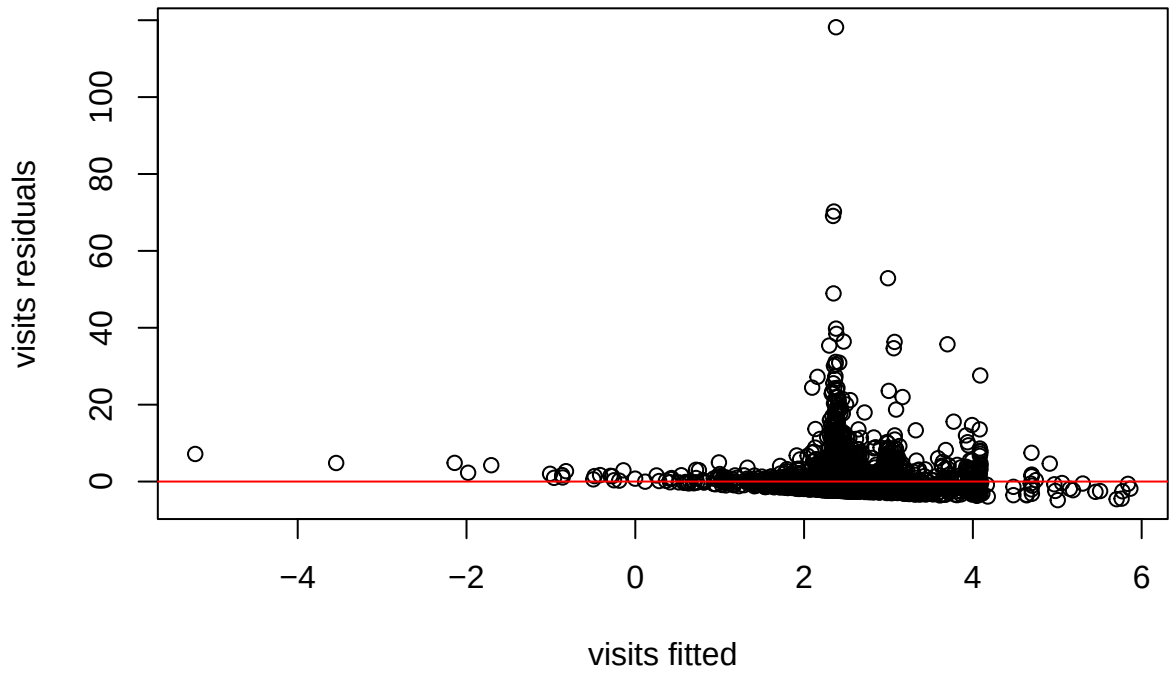
bringing our data from 9315 to 9260.

In order to approach regression in a different way, we also used ANOVA in order to assess the different dependent variables. We wanted to determine if libraries that have significantly different visiting levels, user levels, and circulation levels. We assigned each library a level 1-5 based on it's quantile position. For example, if it is under the median visits but above the 2nd quantile visits, it would be labeled as 2.

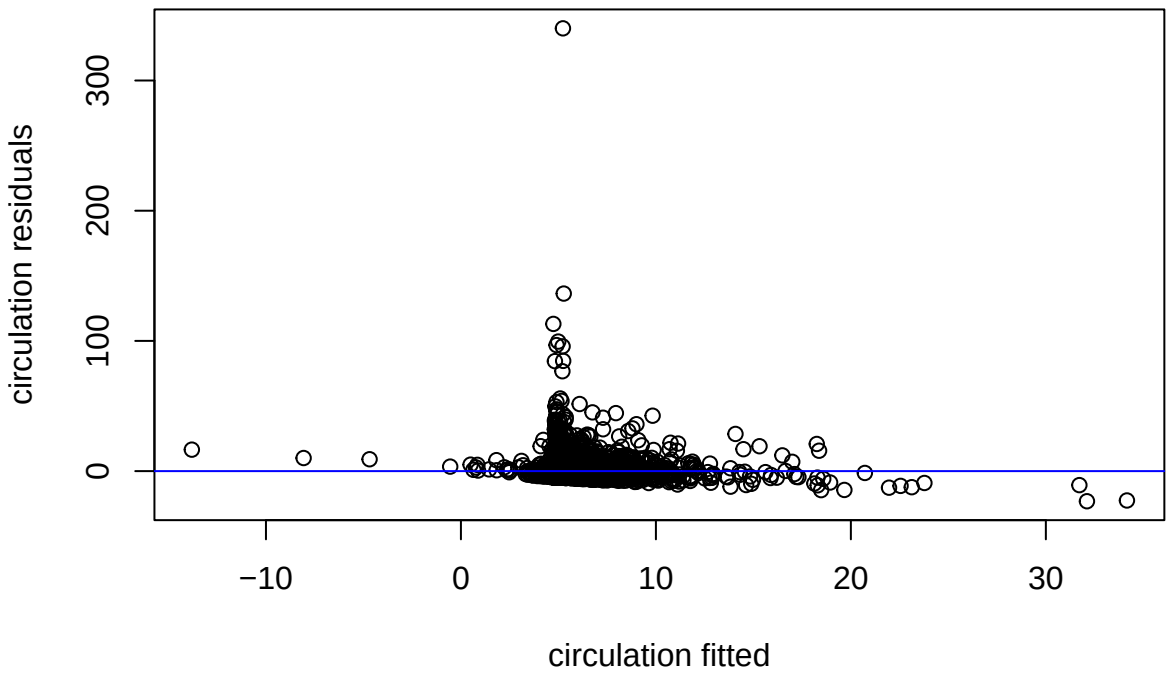
For each of the dependent variables, we used the formula LOANTO~dependent variable. We chose LOANTO because it is a variable that represents the activity & influence of a library without being directly related to any of the three variables. We removed the same entries as the linear regression

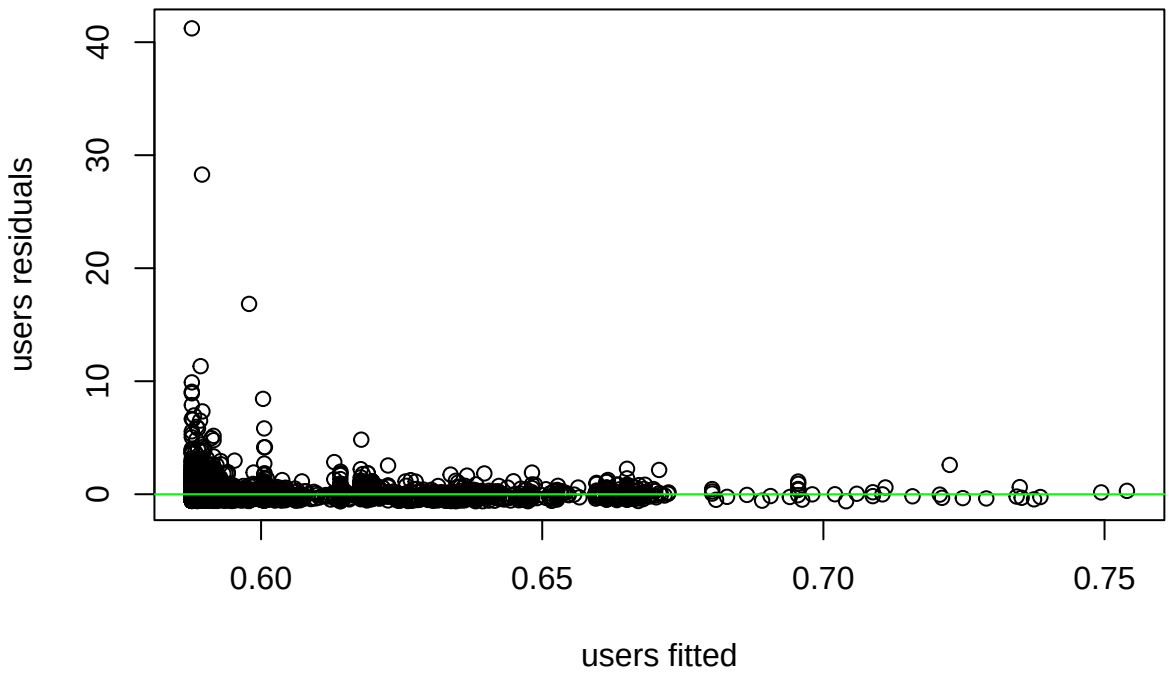
3.2.1 For linear regression: The significant variables for Visits were: EBOOK + PRMATEXP + VIDEO_DL + PHYSCIR + VIDEO_PH. The significant variables for Circulation were: EBOOK + CAPITAL + PRMATEXP + PITUSR + LOANTO + ELMATCIR + PHYSCIR The significant variables for Users were:EBOOK

The following plots demonstrate the relationship between the fitted values & the residuals for these three mod-



els:

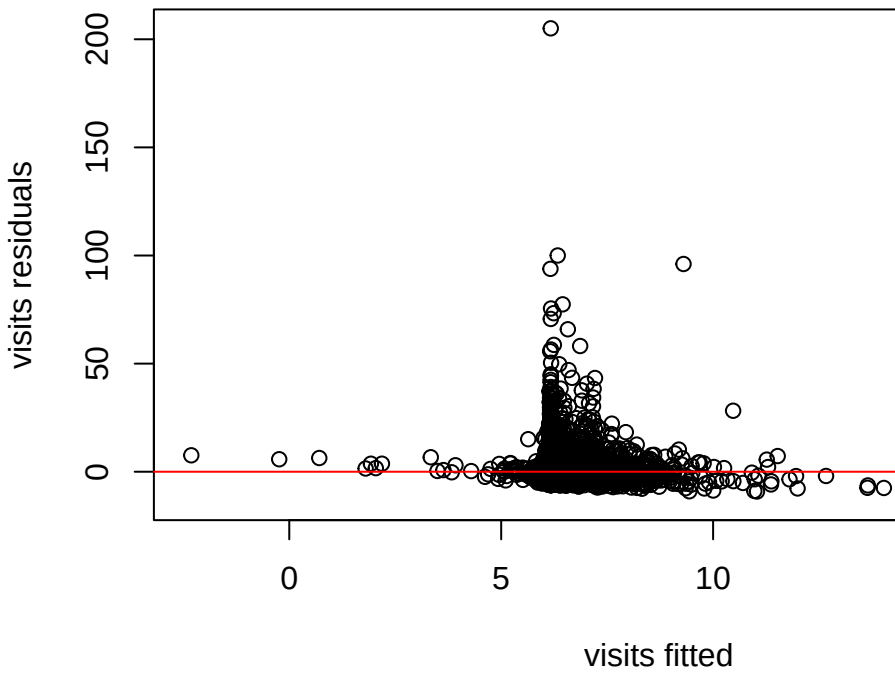

circulation residuals
0
100
200
300
−10
0
10
20
30
circulation fitted



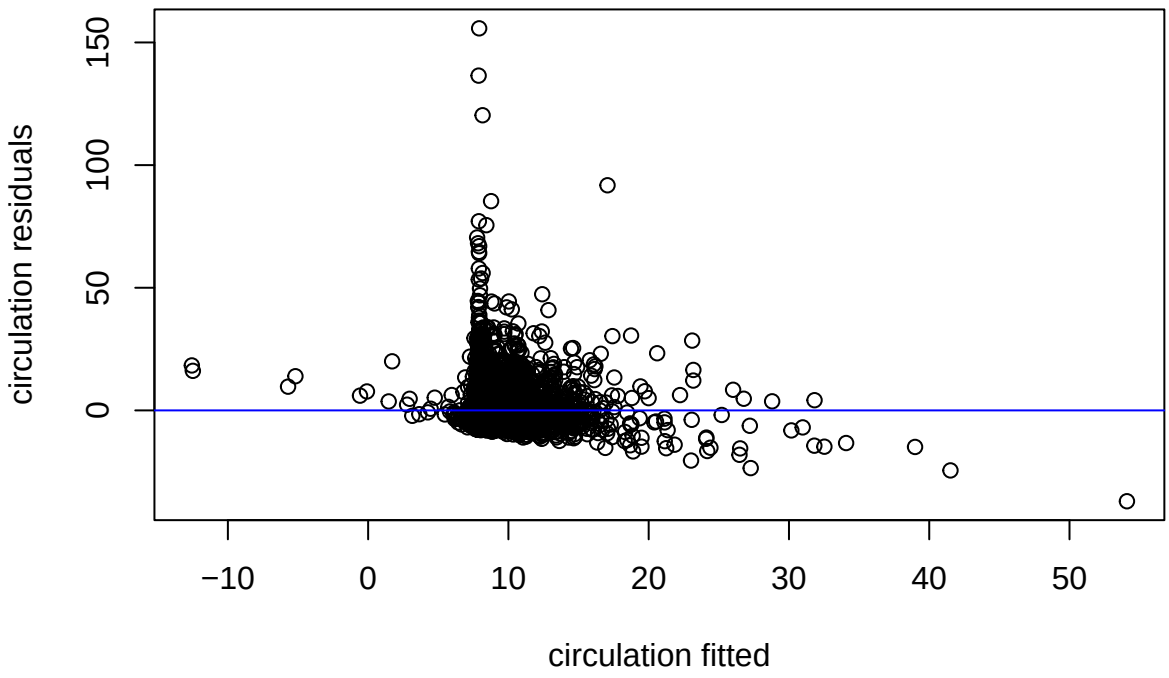
As you can see, the residuals are not well distributed across all fitted values. This implies that there are more features that must be considered.

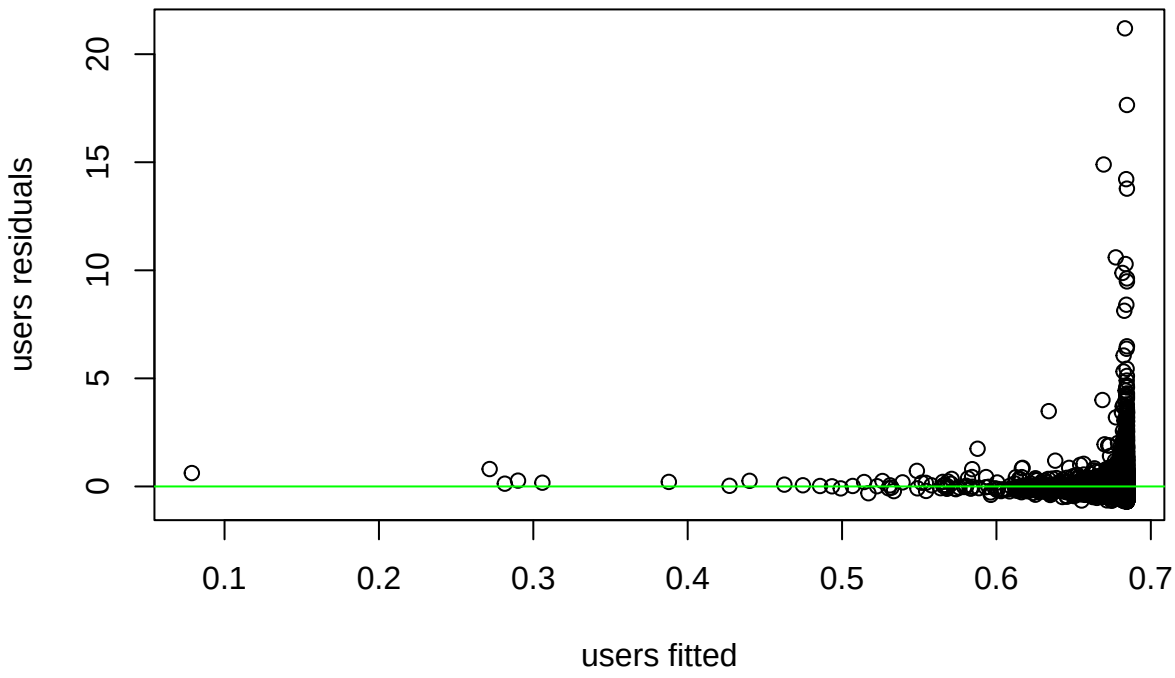
When we fit a linear model to 2011, the significant variables change, but the residuals are still not evenly distributed. In 2011:

The significant variables for Visits were: BKVOL + EBOOK + LOANTO + AUDIO_DL The significant variables for Circulation were: BKVOL + EBOOK + PRMATEXP + PITUSR + LOANTO + LOANFM + AUDIO_DL + VIDEO_DL The significant variables for Users were:LOANFM



The following residual plots for 2011 are below:


circulation residuals
150
100
50
0
−10
0
10
20
30
40
50
circulation fitted

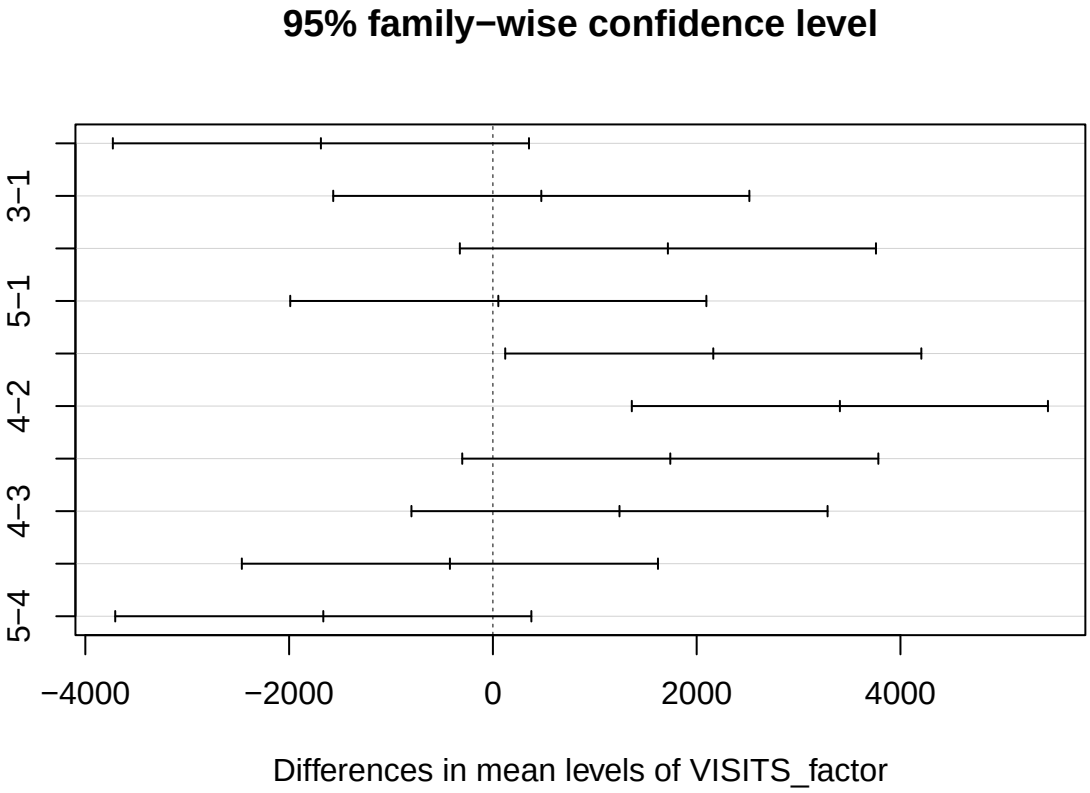


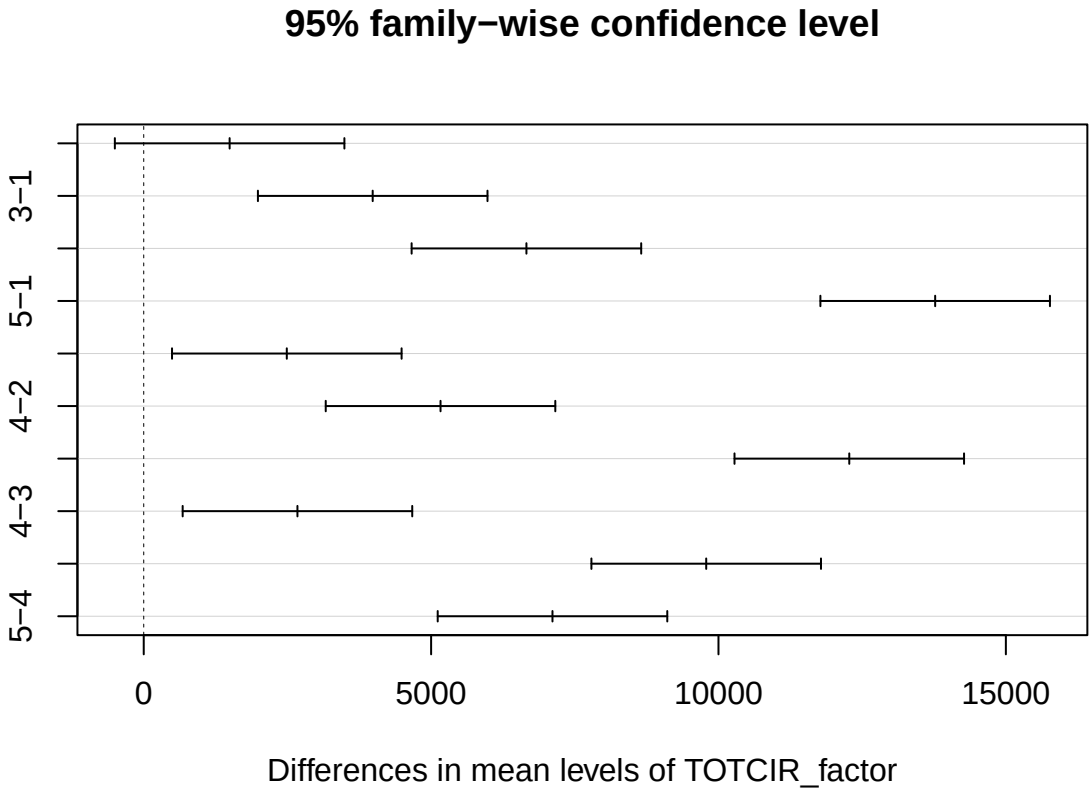
####3.2.2 For ANOVA, the now-independent dependent variables have significant differences between the quantile groups for all 3 dependent variables.

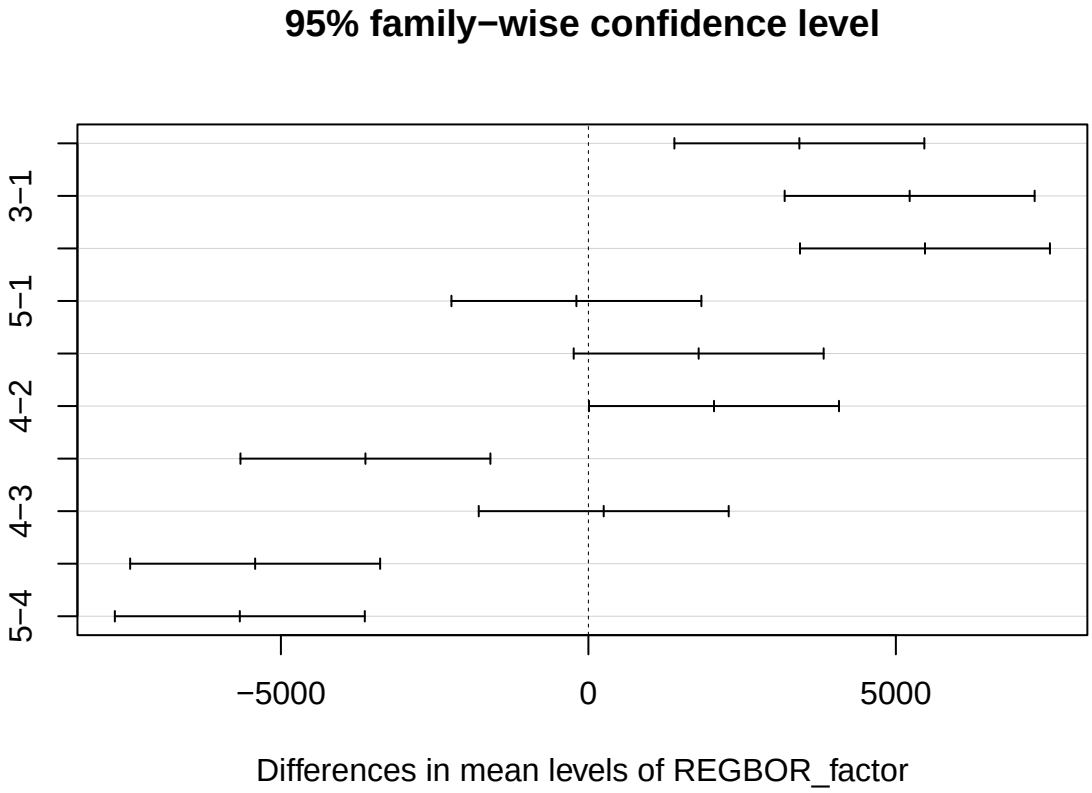
```
##     LIBID              LIBNAME             STABR              C_LEGBAS
##  Length:9215        Length:9215        Length:9215        CI     :4850
##  Class :character   Class :character   Class :character   LD     :1405
##  Mode  :character   Mode  :character   Mode  :character   NP     :1328
##                                                           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      PITUSR            LOANTO             LOANFM
##  0 :9140   Min.   :     -3   Min.   :    -3.0   Min.   :    -3
##  6 :  20   1st Qu.:    306   1st Qu.:    31.5   1st Qu.:    50
##  14:  55   Median :   1090   Median :   548.0   Median :   593
##            Mean   :   5751   Mean   :  6969.2   Mean   :  7153
##            3rd Qu.:   3610   3rd Qu.:  5436.5   3rd Qu.:  4832
##            Max.   :1484987   Max.   :706516.0   Max.   :919237
##
##      TOTCIR             REGBOR           AUDIO_PH         AUDIO_DL
##  Min.   :      -3   Min.   :     -3   Min.   :    -3   Min.   :      -3
##  1st Qu.:    7546   1st Qu.:    990   1st Qu.:   316   1st Qu.:    5003
##  Median :   25986   Median :   3136   Median :  1041   Median :   16032
##  Mean   :  168778   Mean   :  17293   Mean   :  4072   Mean   :   55286
##  3rd Qu.:   90500   3rd Qu.:  10131   3rd Qu.:  3314   3rd Qu.:   34278
##  Max.   :17408320   Max.   :2696713   Max.   :511890   Max.   :16457956
##
##     VIDEO_PH         VIDEO_DL
##  Min.   :    -3   Min.   :    -3
##  1st Qu.:  1080   1st Qu.:     0
##  Median :  2490   Median :   104
##  Mean   :  7221   Mean   :  4861
##  3rd Qu.:  6006   3rd Qu.:  1169
##  Max.   :525571   Max.   :663182
##
```

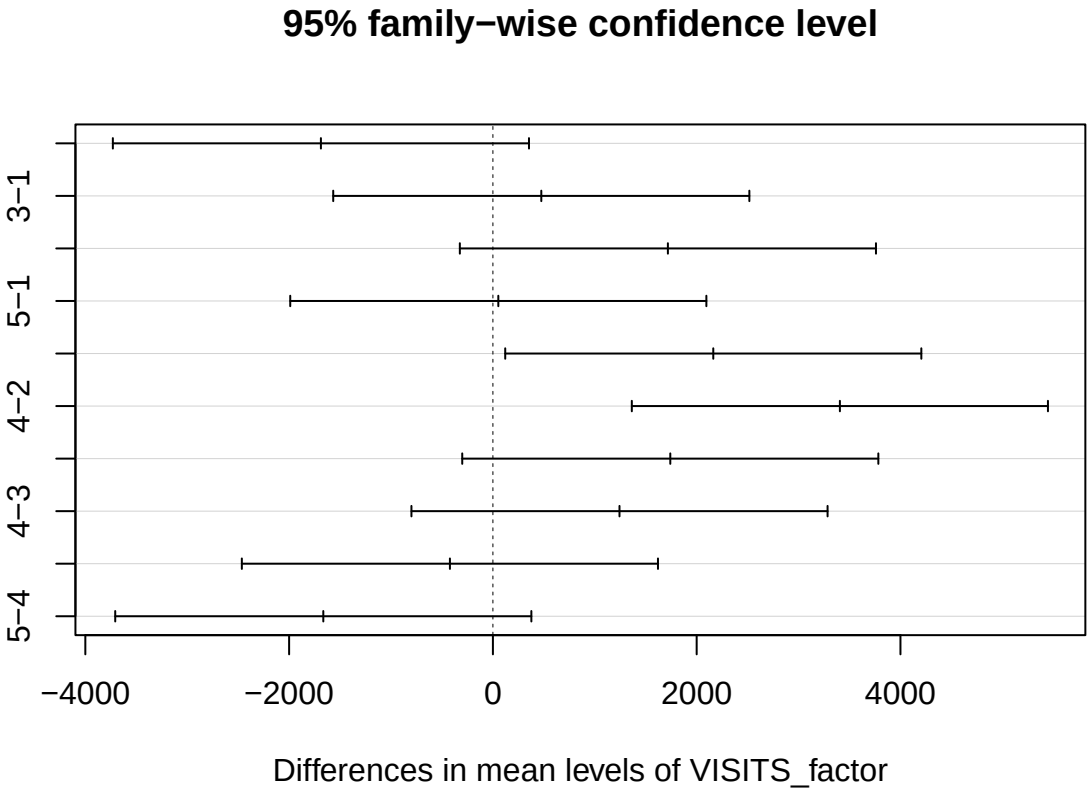
```
##        LIBID                       LIBNAME STABR C_LEGBAS POPU_LSA
## 1 AK0001-002  ANCHOR POINT PUBLIC LIBRARY    AK       NP     2123
## 2 AK0002-011     ANCHORAGE PUBLIC LIBRARY    AK       CO   288970
## 3 AK0003-002   ANDERSON COMMUNITY LIBRARY    AK       CI      275
## 4 AK0006-002 KUSKOKWIM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

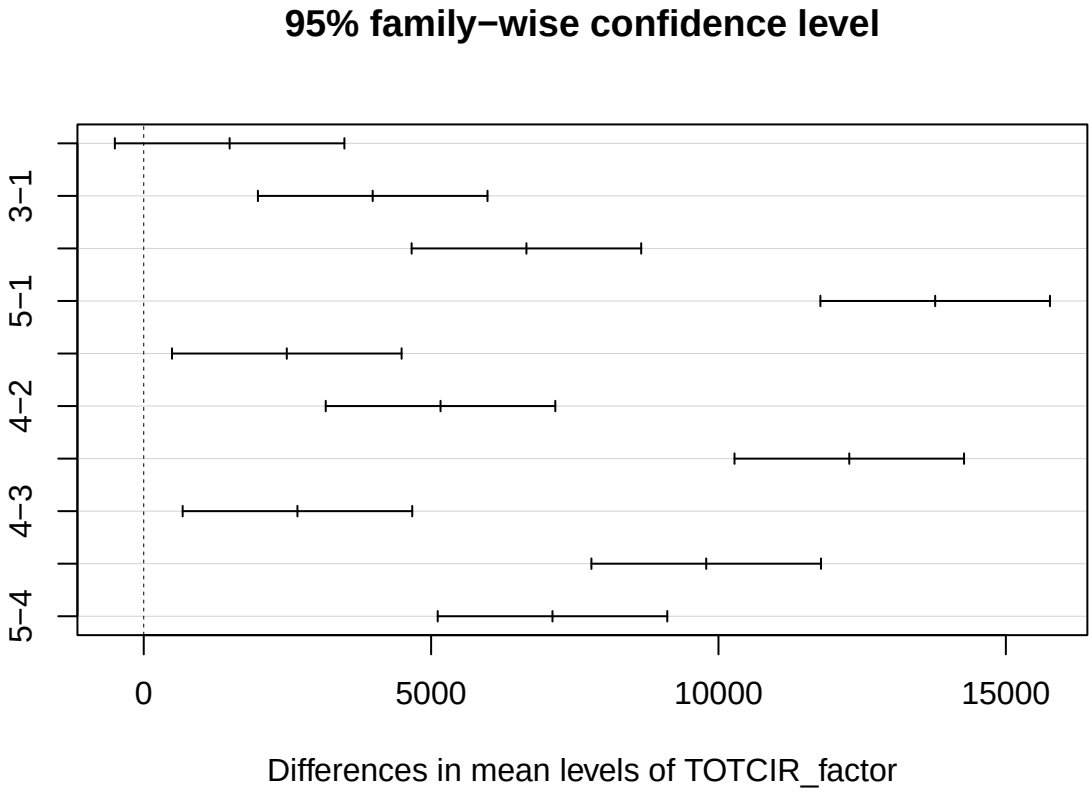

95% family−wise confidence level
3−1
5−1
4−2
4−3
5−4
−4000
−2000
0
2000
4000
Differences in mean levels of VISITS_factor

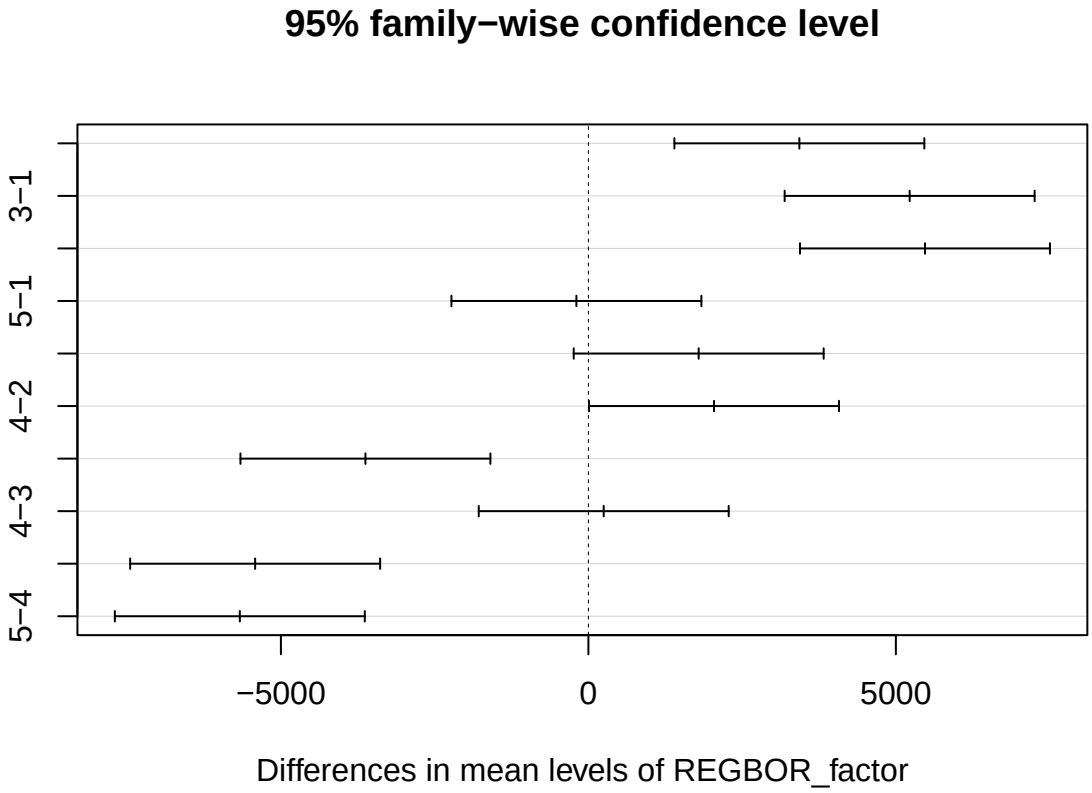

95% family−wise confidence level
3−1
5−1
4−2
4−3
5−4
0
5000
10000
15000
Differences in mean levels of TOTCIR_factor



The same ANOVA formulas for the 2011 data result in similar differences. As with 2021, 2011 also shows significant differences in means for all three dependent variables.


95% family−wise confidence level
3−1
5−1
4−2
4−3
5−4
−4000
−2000
0
2000
4000
Differences in mean levels of VISITS_factor


95% family−wise confidence level
3−1
5−1
4−2
4−3
5−4
0
5000
10000
15000
Differences in mean levels of TOTCIR_factor



4.0 Conclusions Going forward, we are planning to compare more of the years in a time-series type of way. We were surprised to see patterns in the residuals of both the 2011 data and the 2021 data. We plan to explore more variables & their effect on the linear models in order to address this. We weren't surprised to see that different quantile groups of the dependent variables had different LOANTO means.

With the clustering, we were surprised to see how significant the outliers were & how many had to be removed in order to properly model the data. Our hypothesis is that there are significant outliers for certain types of libraries such as high tourist locations or historical landmarks.

As for our research questions, the different significant variables for the linear regression between the years implies that impactful features have changed over the years. The disappearance of the number of regular books as a significant factor is particularly interesting & prompts further inquiry.

With clustering, since region doesn't significantly impact our results based on the results of the clustering, we want to focus more on the time evolution aspect of libraries rather the geographical relevancy aspect of libraries.

5.0 References

https://forum.posit.co/t/i-am-trying-to-remove-rows-that-have-a-certain-value-in-a-specific-column-but-am-getting-an-error/166917/2 https://andrea-grianti.medium.com/kmeans-parameters-in-rstudio-explained-c493ec5a05df https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/