CSC 367 Final Project Report

Project: Public Library Data Set

Reporter: Yingping Li

Table of Contents

Abstract

Introduction

Methods

Discussion & Results

- **CFA**
- **PCA**
- **Clustering**
- **Decision tree**

Conclusion

Reference & Appendix

Abstract

The purpose of this report is to provide our readers an insight of which state has the most books per capita. The data set being used is the 2014 Public Libraries Survey (PLS) data set, which can be found on Kaggle. The nature of the problem involves with two major areas - library resources and financials. First, the data was imported into R where it was preprocessed and cleaned. Once ready, four techniques for the analysis - common factor analysis (CFA), principal component analysis (PCA), clustering analysis, and decision tree were used. It was found that there are significant correlations between the variables pertaining to expenditures and revenues. There are also some significant relationships between library visits and the types of media that location holds in its collection. Moreover, the library resources of the 51 states can be classified into three categories based on certain predictors.

Introduction

About the Data Set

The Public Libraries Survey (PLS) is performed annually by the Institute of Museum and Library Services under the mandate in the Museum and Library Services Act of 2010. The data file includes all public libraries identified by state library administrative agencies in the 50 states and the District of Columbia. The reporting unit for the survey is the administrative entity, defined as the agency that is legally established under local or state law to provide public library service to the population of a local jurisdiction. The FY 2014 PLS collected state characteristics data, including the state total population estimate, number of central and branch libraries, total library visits, circulation transactions, and data on each public library such as its name, location, population of legal service area, print and digital collections, full-time-equivalent staff, and operating revenue and expenditures.

Data Preprocessing

Public Library Data Set has two data sets. One is based on library cases, and the other is based one state cases. The raw table 1 contains 9242 rows and 74 columns and the raw table 2 contains 51 rows and 62 columns. The objectives of libraries include increasing the number of registered users, revenue, and total visits. Because there are so many variables present in the initial data set, I examined it to see if anything could be done to better my final analysis results. First, I remove unrelated and standard columns such as the start and end date of when the data was collected. Additionally, some columns have more missing values than valid values, which are neither appropriate for imputation nor helpful for our analysis. I decided to simply remove such columns. The remaining data set now contains 56 columns. I replaced missing data with mean when the data is normally distributed and replaced missing data with medium when the data is not normally distributed. I can now see that there are some patterns in the data set. Some columns are about expenditures and revenues, whereas other columns concern different library collections. These patterns will be helpful for our analyses.

Methods

In this report, four methods: common factor analysis (CFA), principal component analysis (PCA), clustering analysis, and decision tree were utilized.

I want to use factor analysis to reveal underlying factors in the survey data. CFA seeks the least number of factors which can account for the common variance (correlation) of a set of variables. It can reduce the number of variables needed to run a model, while also providing some insights into the underlying relationships among the data.

Principal Component Analysis (PCA) can extract the important information from a multivariate data set and express this information as a new parsimonious set of variables called principal components. These new variables correspond to linear combinations of the original set. There will only be fewer principal components than number of variables in the original set. We can use PCA to find new, underlying relationships between the variables that could not be understood from the raw data. Interpreting the results of the model will show us these relationships between the variables in each component.

For the clustering analysis, principal component analysis was first performed on the variables to reduce the dimensionality of the attributes. With too many variables in the data set, the clusters would become unwieldy and potentially difficult to interpret properly. Nine variables were created from the data table 2 to generate three and four clusters with the k-means method of clustering. The results are shown in the graphs below.

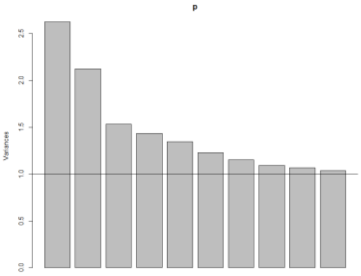
Decision tree analysis was used to classify the statues of the library resources of the 51 states. Decision tree analysis is a non-parametric supervised learning method and can be used for classification and regression. The purpose is to build a model that predicts the value of a target variable, the clusters from the k-means clustering analysis by learning simple decision rules implied from the data features.

Discussion and Results

CFA

CFA is common factor analysis, the point of using this method is to find any shared variance within the data. I want to use factor analysis to reveal underlying factors in the survey data. CFA tends to be used on large amounts of data to uncover latent factors in the data and any underlying relationships. I used 32 variables in my common factor analysis. I wanted to see if there was any relation between those variables or if they would reveal any underlying factors. The knee of the Screen Plot below suggests that three components can be used to represent variables with more shared information.

Figure 1: Screen Plot



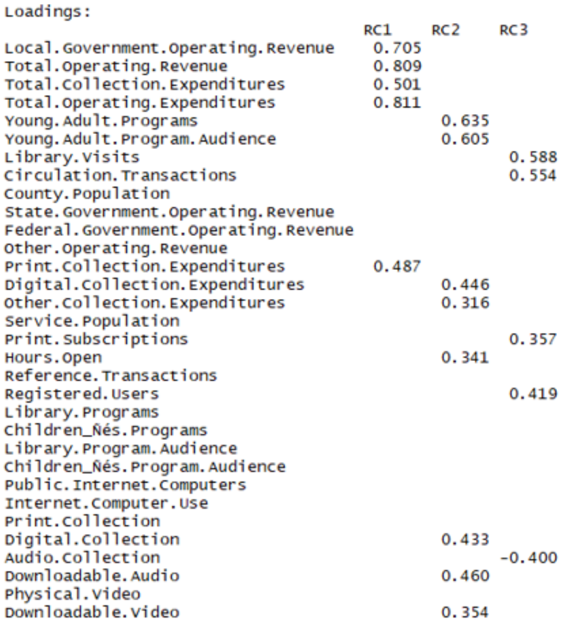
PCA

I want to use PCA to find the relationships between the variables. After I removed some variables which have too much missing values which is greater than 20 percent, 32 variables were selected and checked correlations as result of without strong multicollinearity between each variable. Then 32 variables were applied by the PCA for dimension reduce and CFA for underlying factors. After I compared the two methods' results, I found there are not much difference between them. Also, I compared the results of three and four components and decided using three components RC1, RC2, and RC3 which is in line with the CFA screen plot result.

From the PCA outputs in Table 1, I named the three components RC1, RC2, and RC3 respectively as "Revenue", "Program", and "Visit", and I created Table 2 which shows the variables and loadings of each component.

The Revenue component includes variables: Local Government Operating Revenue, Total Operating Revenue, Total Collection Expenditures, Total Operating Expenditures, Print Collection Expenditures with loading as 0.705, 0.809, 0.501, 0.811, and 0.487, respectively. Program component includes variables: Young Adult Programs, Young Adult Program Audience, Digital Collection Expenditures, Other Collection Expenditures, Hours Open, Digital Collection, Downloadable Audio, and Downloadable video with loading as 0.635, 0.605, 0.446, 0.316, 0.433, 0.460, and 0.354 respectively. Visit component includes variable: Library Visits, Circulation Transactions, Print Subscriptions, Registered Users, Audio Collection with loading 0.588, 0.554, 0.357, and - 04, respectively

Table 1: PCA outputs



```
Loadings:
                                     RC1    RC2    RC3
Local.Government.Operating.Revenue    0.705
Total.Operating.Revenue               0.809
Total.Collection.Expenditures         0.501
Total.Operating.Expenditures          0.811
Young.Adult.Programs                         0.635
Young.Adult.Program.Audience                 0.605
Library.Visits                                      0.588
Circulation.Transactions                            0.554
County.Population
State.Government.Operating.Revenue
Federal.Government.Operating.Revenue
Other.Operating.Revenue
Print.Collection.Expenditures         0.487
Digital.Collection.Expenditures              0.446
Other.Collection.Expenditures                0.316
Service.Population
Print.Subscriptions                                 0.357
Hours.Open                                   0.341
Reference.Transactions
Registered.Users                                    0.419
Library.Programs
Children_Ñés.Programs
Library.Program.Audience
Children_Ñés.Program.Audience
Public.Internet.Computers
Internet.Computer.Use
Print.Collection
Digital.Collection                           0.433
Audio.Collection                                   -0.400
Downloadable.Audio                           0.460
Physical.Video
Downloadable.Video                           0.354
```

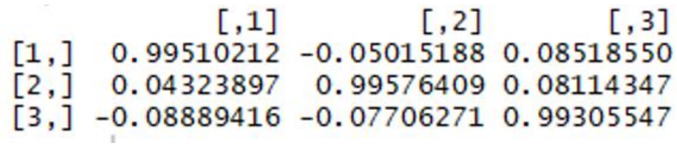
Table 2: PCA 3 components and loading

component	variable	loading
Revenue	Local Government Operating Revenue	0.705
	Total Operating Revenue	0.809
	Total Collection Expenditures	0.501
	Total. Operating. Expenditures	0.811
	Print Collection Expenditures	0.487
Program	Young Adult Programs	0.635
	Young Adult Program Audience	0.605
	Digital Collection Expenditures	0.446
	Other Collection Expenditures	0.316
	Hours Open	0.341
	Digital Collection	0.433
	Downloadable Audio	0.46
	Downloadable video	0.354
Visit	Library Visits	0.588
	Circulation Transactions	0.554
	Print Subscriptions	0.357
	Registered Users	0.419
	Audio Collection	-0.400

.

The diagonal line in the Table 3 shows that there is a strong relationship insidde each component’s variables because the values are greater than 0.9, but there is a weak relationship between different components.

Table 3: PCA three components correlation

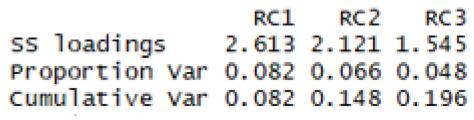


```
              [,1]        [,2]       [,3]
[1,]  0.99510212 -0.05015188 0.08518550
[2,]  0.04323897  0.99576409 0.08114347
[3,] -0.08889416 -0.07706271 0.99305547
```

Table 2 and Table 3 revel some underlying relationships among the variables. For example, Local Government Operating Revenue, Total Operating Revenue, Total Collection Expenditures, Total Operating Expenditures, Print Collection Expenditures strongly relate each other and this can make sense because the expenditures depend on the revenues. Also, the Program component reveals that Digit Collection expenditures, Other Collection Expenditures strongly relate to the Young Adult Programs, Young Adult Program Audience Digital Collection Expenditures, Downloadable Audio, and Downloadable video. Furthermore, the visit component shows there is a negative loading (-0.4) on Audio Collection which means that there is an inverse relationship between the Audio Collection and other variables including Library Visits, Print Subscriptions, and Registered Users. This may reveal a trend in which readers can use Audio Collection instead of visiting the library and using prints collections.

From the Figure 2, I find three components with 20 percent cumulative variances which is not very high but for the survey program it still could be normal.

Figure2: PCA three components cumulative variance



```
                 RC1   RC2   RC3
SS loadings    2.613 2.121 1.545
Proportion Var 0.082 0.066 0.048
Cumulative Var 0.082 0.148 0.196
```

<u>Clustering</u>

For clustering, we would like to see if there are any underlying patterns in the data based on exploratory grouping of the data entries and similarities in their attributes.

I applied K-means clustering and Hierarchical clustering on data table 1 which has and 9242 rows and 56 columns. The k-means clustering's outputs in Figure 3 shows that three clusters can mainly represent the data because the knee of the curve is around 3. Also, Hierarchical clustering outputs in Figure 4 shows that three clustering can represent the data. These results are in line with the result of PCA.

Figure 3: K-means clustering

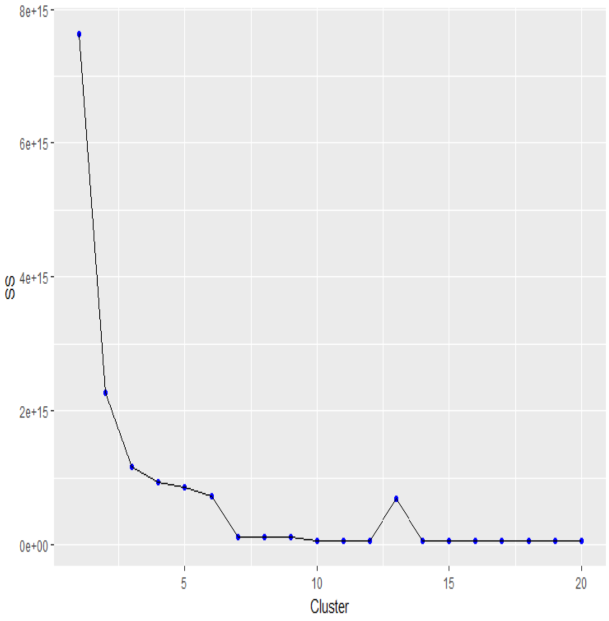
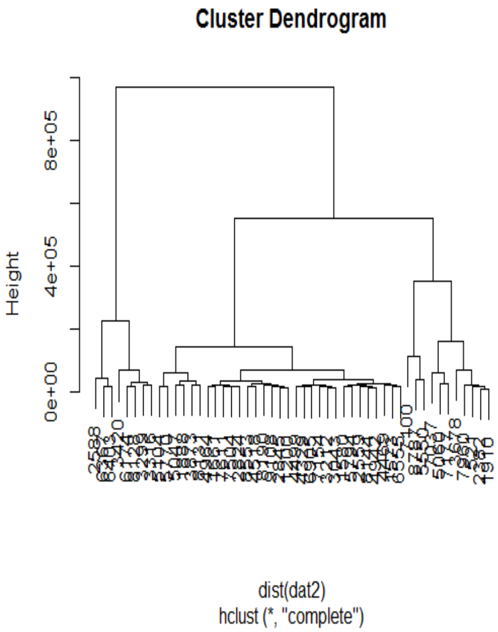


Figure 4: Hierarchical clustering



After I applied CFA, PCA, clustering analysis on the data table 1, I obtained some information about the data and decided to find the relationship between library resources and total operating revenue based on state level. I used the dataset table 2 in which I picked nine variables and divided them by variable --State population. I obtained nine new variables including Total Operating Revenue Per 1000, Library Visits per 1000, Physical Video per 1000, Audio Collection per1000, Library Programs per 1000, Print Collection per 1000, Digital Collection per 1000, Downloadable Audio per 1000, Downloadable Video per 1000. I applied k-means clustering with k=3 and k=4 on these new variables. The outputs of cluster center and number of cases in each cluster shows in Figure 5 and Figure 6 and Figure 7 when k is 3.

Figure5: cluster center of k-means clustering with k=3

Final Cluster Centers

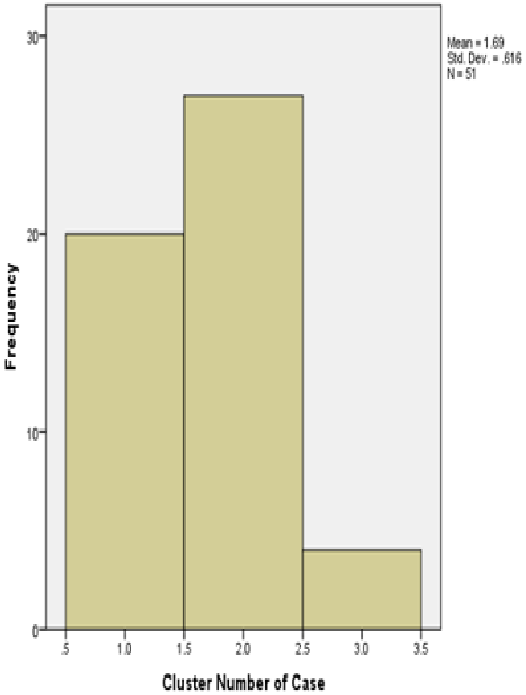
	Cluster		
	1	2	3
Print Collection per 1000	3208.82	2300.69	3374.85
Digital Collection per 1000	1103.35	514.61	1270.14
Audio Collection per1000	189.80	119.59	206.52
Downloadable Audio per 1000	21.62	8.81	28.83
Physical Video per 1000	264.28	159.93	298.01
Downloadable Video per 1000	21.62	8.81	28.83
Library Visits per 1000	5353.57	3981.62	6177.32
Library Programs per 1000	19.44	12.98	21.54
Total Operating Revenue per 1000	47273.28	26690.22	69879.07

Figure 6: number of cases clustering with k=3

Number of Cases in each Cluster

Cluster	1	20.000
	2	27.000
	3	4.000
Valid		51.000
Missing		.000

Figure 7: Histogram of clustering number of cases with k=3



The outputs of the final cluster center and number of cases in each cluster shows in Figure 8 and Figure 9 and Figure 10 when k is 4.

Figure8: cluster center of k-means clustering with k=4

Final Cluster Centers

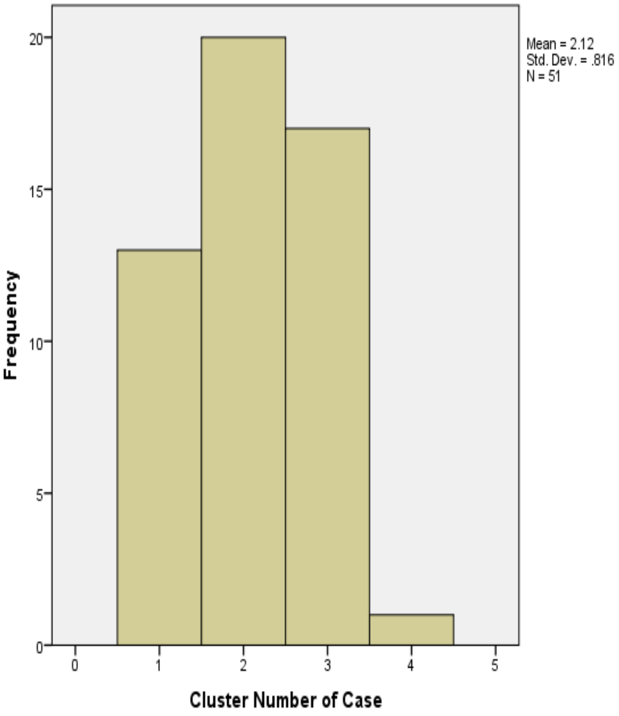
	Cluster			
	1	2	3	4
Print Collection per 1000	3117.11	2119.82	3183.53	2755.44
Digital Collection per 1000	691.77	604.43	1158.33	268.76
Audio Collection per1000	202.21	106.04	174.73	131.15
Downloadable Audio per 1000	16.72	10.20	21.37	1.22
Physical Video per 1000	284.53	141.73	238.17	213.49
Downloadable Video per 1000	16.72	10.20	21.37	1.22
Library Visits per 1000	5563.29	3615.23	5190.34	6421.06
Library Programs per 1000	19.54	11.13	19.24	21.79
Total Operating Revenue per 1000	56175.08	24192.82	38190.78	82242.17

Figure 9: number of cases of k-means clustering with k=4

Number of Cases in each Cluster

Cluster	1	13.000
	2	20.000
	3	17.000
	4	1.000
Valid		51.000
Missing		.000

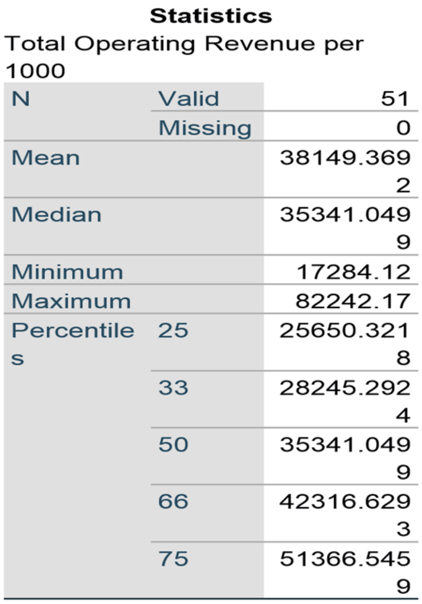
Figure 10: Histogram of clustering number of cases with k=4



<u>Decision tree</u>

I want to use decision tree to classify the library resources, activities, and revenue under the state level. First, I transferred my dependent variable Total Operating Revenue from continuous variable to categorical variable with 3 classes (low, medium, and high) by using equal frequency binning to get balanced data. Figure 11 shows the statistics of the dependent variable which is not normally distributed and without outliers. The Total Operating Revenue per 1000 capita ranges from 17284.12 to 82242.17 which represent state MS and DC.

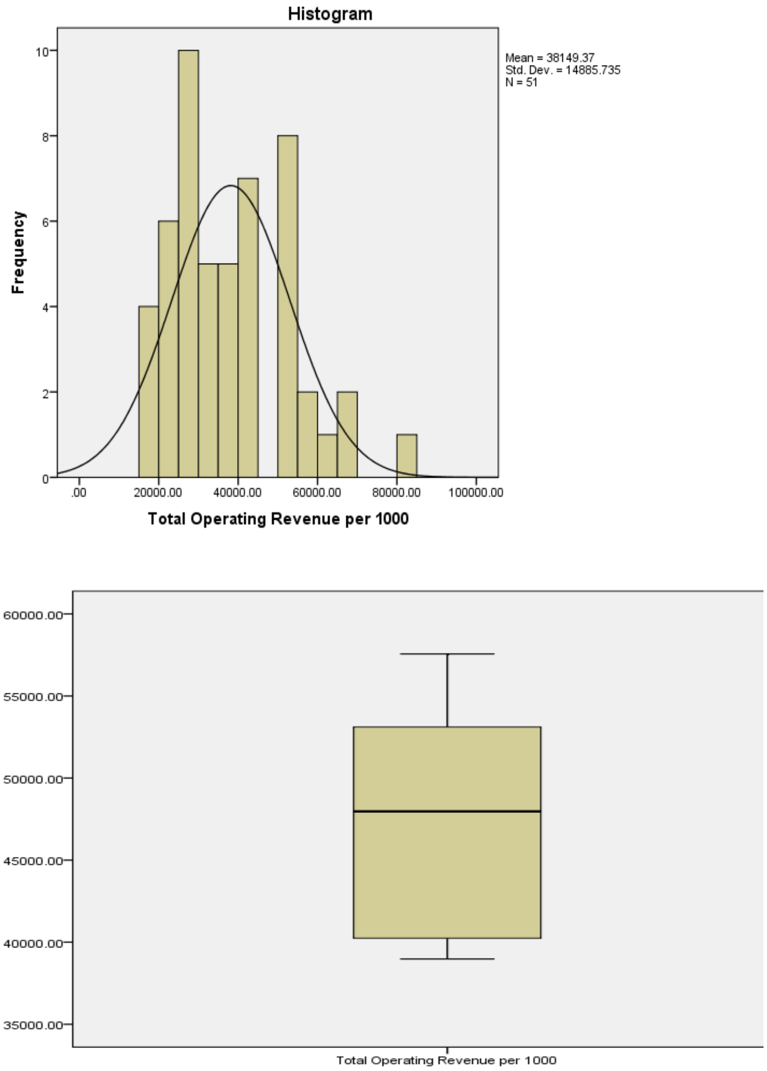
Figure 11: Dependent variable statistic for decision tree 1



Statistics

Total Operating Revenue per 1000

N	Valid	51
	Missing	0
Mean		38149.3692
Median		35341.0499
Minimum		17284.12
Maximum		82242.17
Percentiles	25	25650.3218
	33	28245.2924
	50	35341.0499
	66	42316.6293
	75	51366.5459



Then I employed decision tree analysis with CRT (Classification and Regression Tree). I created the decision tree 1, and the outputs show in the tables below. In Figure 12, the Classification table shows that the model accuracy is 86.3%, but the class 2 only has 79.2% accuracy. In Figure 13, the Independent variable importance table shows that Library Visits per 1000, Physical Video per 1000, Audio Collection per1000 are three most important variables for the model. Figure 14 shows that the tree has nine terminal nodes, and its depth is five.

Figure 12: Decision tree 1 Model Summary and Classification accuracy

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	TotalOperatingRevenuePerCat2
	Independent Variables	Print Collection per 1000, Digital Collection per 1000, Audio Collection per1000, Downloadable Audio per 1000, Physical Video per 1000, Downloadable Video per 1000, Library Visits per 1000, Library Programs per 1000
	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	7
	Minimum Cases in Child Node	3
Results	Independent Variables Included	Library Visits per 1000, Physical Video per 1000, Audio Collection per1000, Library Programs per 1000, Print Collection per 1000, Digital Collection per 1000, Downloadable Audio per 1000, Downloadable Video per 1000
	Number of Nodes	11
	Number of Terminal Nodes	6
	Depth	5

Classification

	Predicted			
Observed	1.00	2.00	3.00	Percent Correct
1.00	13	0	0	100.0%
2.00	4	19	1	79.2%
3.00	1	1	12	85.7%
Overall Percentage	35.3%	39.2%	25.5%	86.3%

Growing Method: CRT
Dependent Variable: TotalOperatingRevenuePerCat2

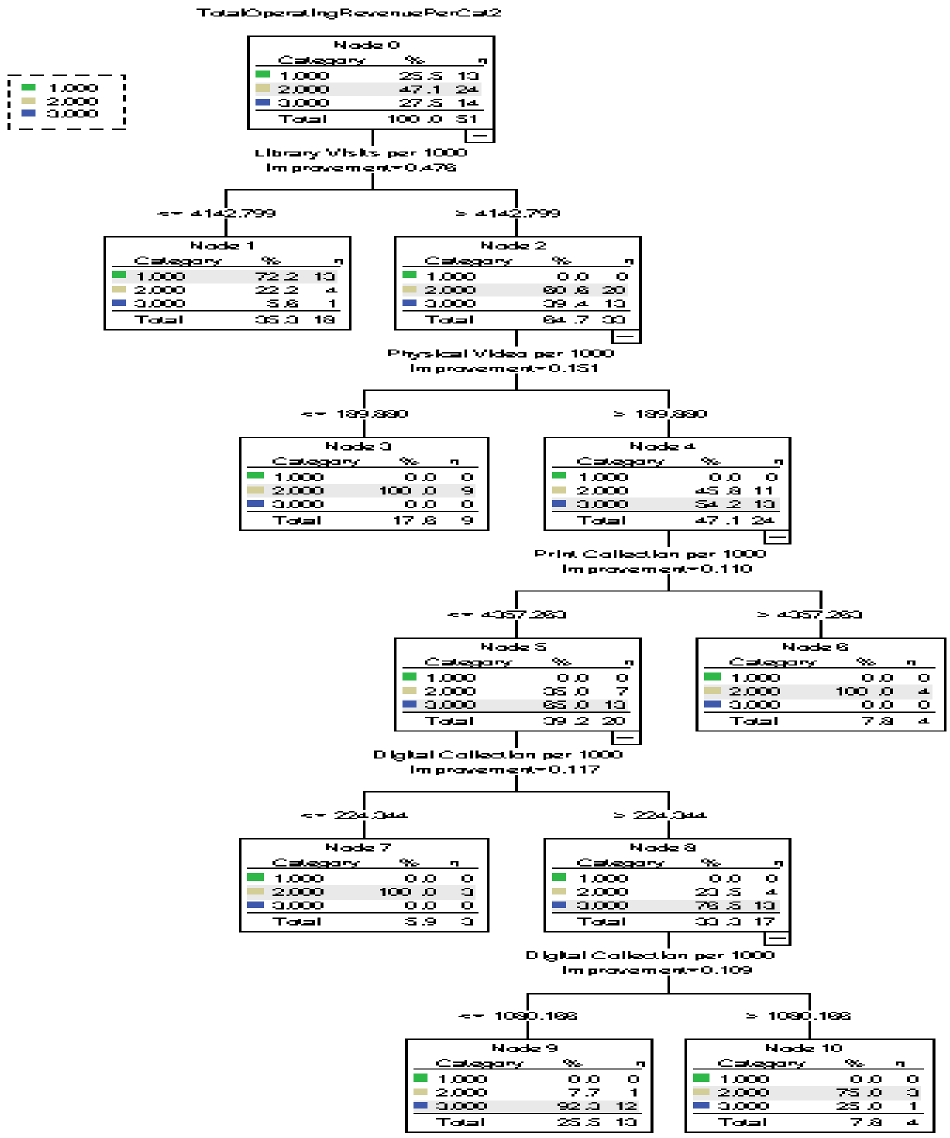
Figure 13: Decision tree 1 variable importance

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
Library Visits per 1000	.589	100.0%
Physical Video per 1000	.560	95.1%
Audio Collection per1000	.518	88.0%
Library Programs per 1000	.399	67.7%

Digital Collection per 1000	.344	58.5%
Print Collection per 1000	.288	48.9%
Downloadable Video per 1000	.140	23.7%
Downloadable Audio per 1000	.140	23.7%

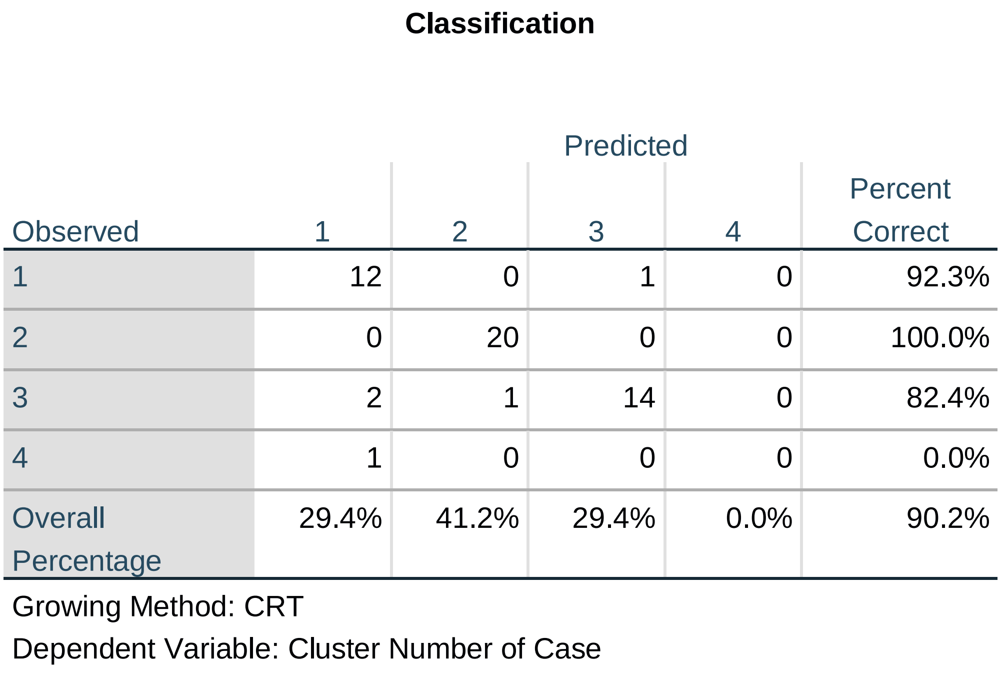
Figure 14: Decision tree1



I want to improve the model's accuracy and I used 4 clusters from clustering analysis as dependent variable to create decision tree 2 under the condition of without Total Operating Revenue and decision tree 3 with Total Operating Revenue. Figure 15 shows the accuracy is

90.2%. Figure 16 shows the accuracy is 98%. This tells me that Total Operating Revenue played an important role in improving the model's accuracy. Node 4 in Figure 16_1 shows in only one case which formed class 4 cannot be correctly classified by decision tree 3.

Figure 15: Decision tree 2 under 4 clusters as dependent variable without Total Operating Revenue per 1000

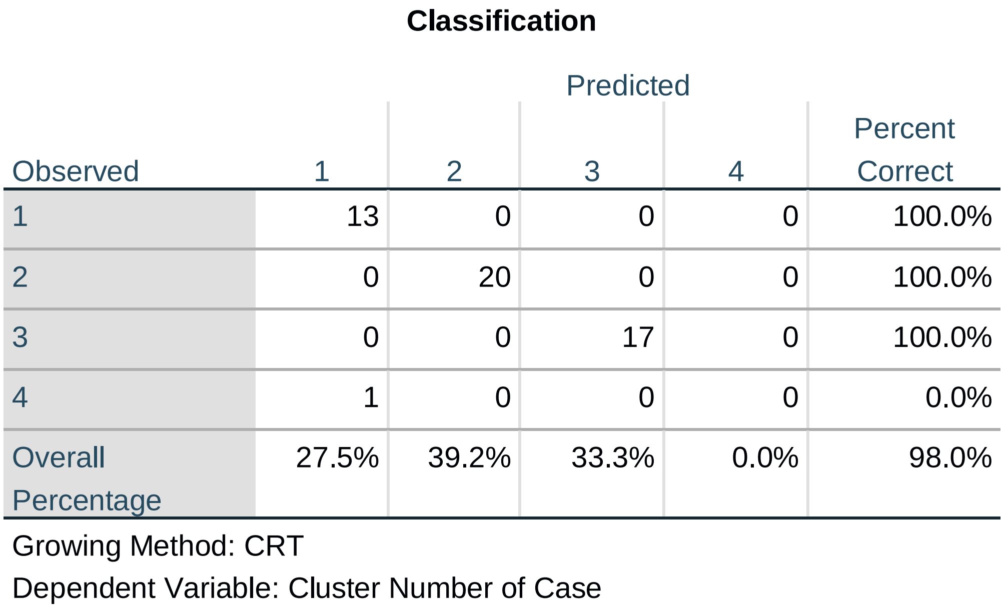


Classification

	Predicted				
Observed	1	2	3	4	Percent Correct
1	12	0	1	0	92.3%
2	0	20	0	0	100.0%
3	2	1	14	0	82.4%
4	1	0	0	0	0.0%
Overall Percentage	29.4%	41.2%	29.4%	0.0%	90.2%

Growing Method: CRT
Dependent Variable: Cluster Number of Case

Figure 16: Decision tree 3 under 4 clusters as dependent variable with Total Operating Revenue per 1000

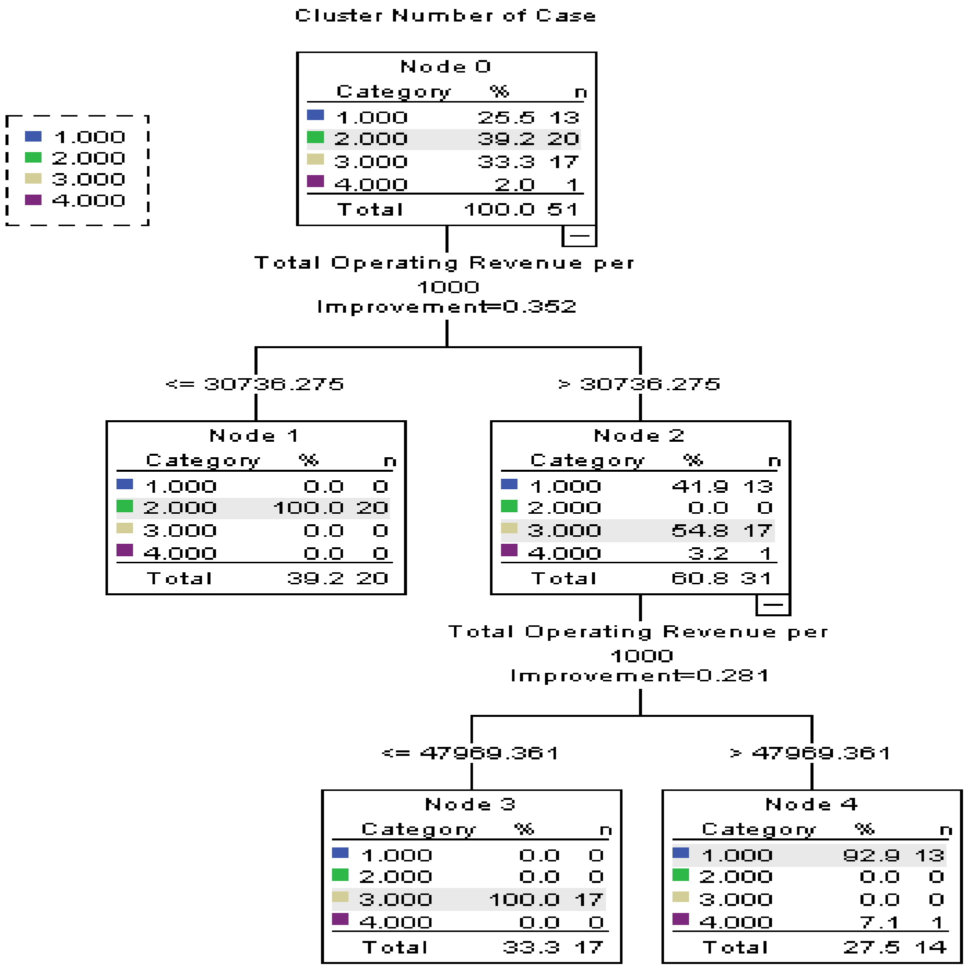


Classification

	Predicted				
Observed	1	2	3	4	Percent Correct
1	13	0	0	0	100.0%
2	0	20	0	0	100.0%
3	0	0	17	0	100.0%
4	1	0	0	0	0.0%
Overall Percentage	27.5%	39.2%	33.3%	0.0%	98.0%

Growing Method: CRT
Dependent Variable: Cluster Number of Case

Figure 16_1: Decision tree 3 under 4 clusters as dependent variable with Total Operating Revenue per 1000



I want to improve the accuracy, and then I used 3 clusters as dependent variable and created decision tree 4 with nine variables.

Figure 17 shows the accuracy is 100%. Figure 19 shows that the decision tree 4 has total five nodes including three terminal nodes, and it also shows that the tree's depth is two. Figure 20 and Figure 21 show that the model's three most important variables are Total Operating Revenue per 1000, Library Visits per 1000, and Physical Video. From Figure 19 and the rules in Figure 22, I found that only one predictor which is Total Operating Revenue variable can predict the class level of the state with 100% accuracy.

Figure 17: Decision tree under 3 clusters as dependent variable including Total Operating Revenue per 1000

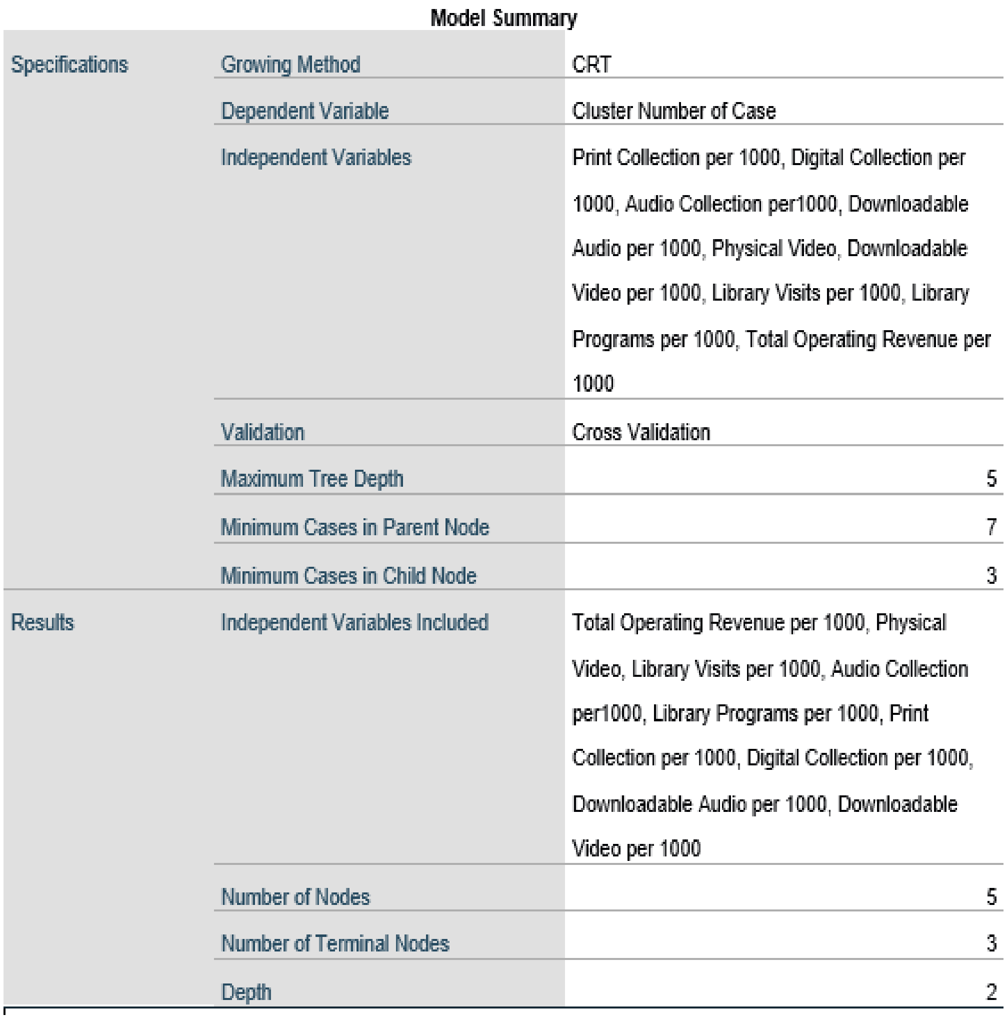
Classification

Observed	Predicted 1	 2	 3	 Percent Correct
1	20	0	0	100.0%
2	0	27	0	100.0%
3	0	0	4	100.0%
Overall Percentage	39.2%	52.9%	7.8%	100.0%

Growing Method: CRT

Dependent Variable: Cluster Number of Case

Figure18: Decision tree 4 under 3 clusters as dependent variable with nine independent variables model summary



Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	Cluster Number of Case
	Independent Variables	Print Collection per 1000, Digital Collection per 1000, Audio Collection per1000, Downloadable Audio per 1000, Physical Video, Downloadable Video per 1000, Library Visits per 1000, Library Programs per 1000, Total Operating Revenue per 1000
	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	7
	Minimum Cases in Child Node	3
Results	Independent Variables Included	Total Operating Revenue per 1000, Physical Video, Library Visits per 1000, Audio Collection per1000, Library Programs per 1000, Print Collection per 1000, Digital Collection per 1000, Downloadable Audio per 1000, Downloadable Video per 1000
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

Figure 19: Decision tree 4 under 3 clusters as dependent variable

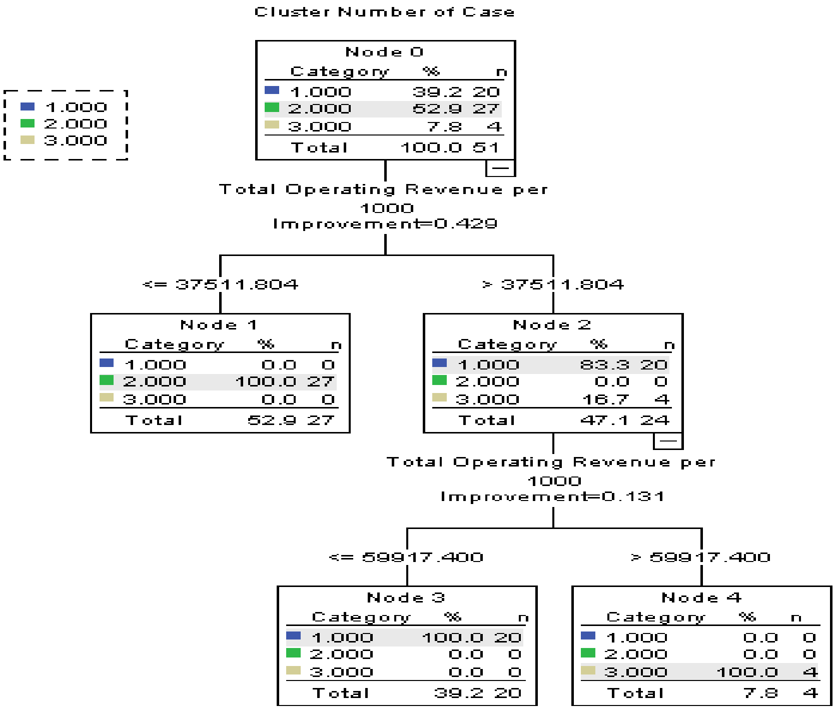
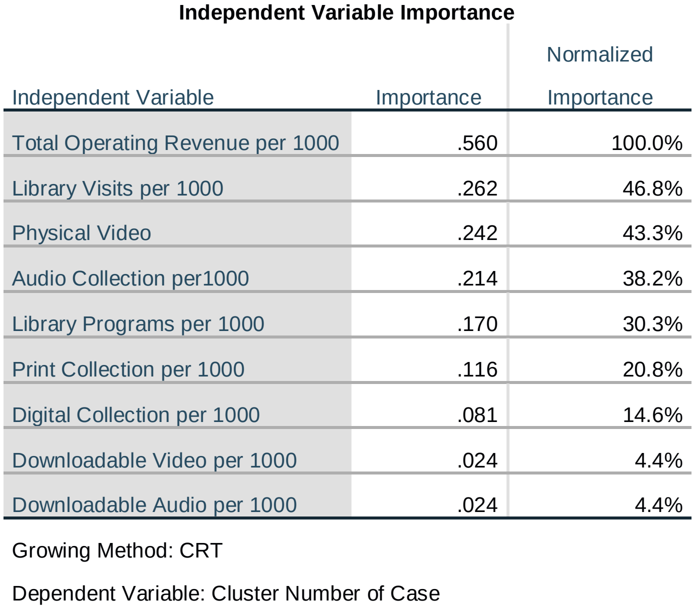


Figure 20: Decision tree 4 under 3 clusters as dependent variable



Independent Variable Importance

Independent Variable	Importance	Normalized Importance
Total Operating Revenue per 1000	.560	100.0%
Library Visits per 1000	.262	46.8%
Physical Video	.242	43.3%
Audio Collection per1000	.214	38.2%
Library Programs per 1000	.170	30.3%
Print Collection per 1000	.116	20.8%
Digital Collection per 1000	.081	14.6%
Downloadable Video per 1000	.024	4.4%
Downloadable Audio per 1000	.024	4.4%

Growing Method: CRT

Dependent Variable: Cluster Number of Case

Figure 21: Decision tree 4 under 3 clusters as dependent variables normalized variable importance

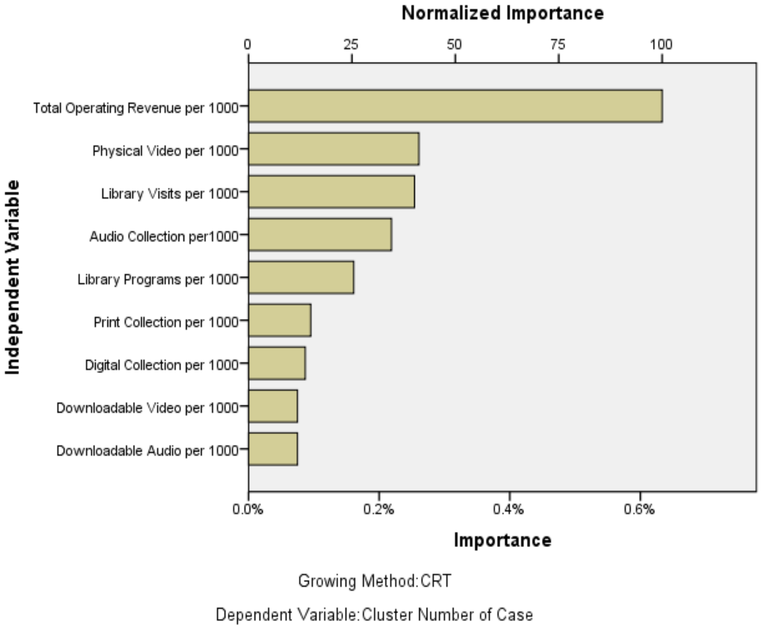


Figure 22: Decision tree 4 rules

```
/* Node 1 */.
DO IF (SYSMIS(TotalOperatingRevenueper1000) OR (VALUE(TotalOperatingRevenueper1000) LE
37511.80442401987)).
COMPUTE nod_001 = 1.
COMPUTE pre_001 = 2.
COMPUTE prb_001 = 1.000000.
END IF.
EXECUTE.

/* Node 3 */.
DO IF (VALUE(TotalOperatingRevenueper1000) GT 37511.80442401987) AND
(SYSMIS(TotalOperatingRevenueper1000) OR (VALUE(TotalOperatingRevenueper1000) LE
59917.39999779017)).
COMPUTE nod_001 = 3.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 1.000000.
END
  IF.
EXECUTE.

/* Node 4 */.
DO IF (VALUE(TotalOperatingRevenueper1000) GT 37511.80442401987) AND
(VALUE(TotalOperatingRevenueper1000) GT 59917.39999779017).
COMPUTE nod_001 = 4.
COMPUTE pre_001 = 3.
COMPUTE prb_001 = 1.000000.
END IF.
EXECUTE.
```

Conclusion

PCA analysis reveals that there is an inverse relationship between the Audio Collection and other variables including Library Visits, Print Subscriptions, and Registered Users. Decision tree 4 model can use one predictor variable- Total Operating Revenue per 1000 capita to classify the level of a state's library resources, visits, and revenue to three classes with 100 percent accuracy. This model tells us which state has the most books per capita. The model classifies that there are twenty states with Total Operating Revenue per 1000 capita lower than 37511.804 and belong to the low class. There are twenty- seven states with Total Operating Revenue per 1000 capita between 37511.804 and 59917.4 and belong to the medium class. There are four states with Total Operating Revenue per 1000 capita higher than 59917.4 and belong to in the high class. The highest Total Operating Revenue per 1000 capita is 82242.17 which is DC (District of Columbia), and the lowest Total Operating Revenue per capita is 17284.12 which is MS (Mississippi). This is in line with the findings reported by the Institute of Museum and Library Services. ("Public Libraries in the United States Survey: Fiscal Year 2014" 17)

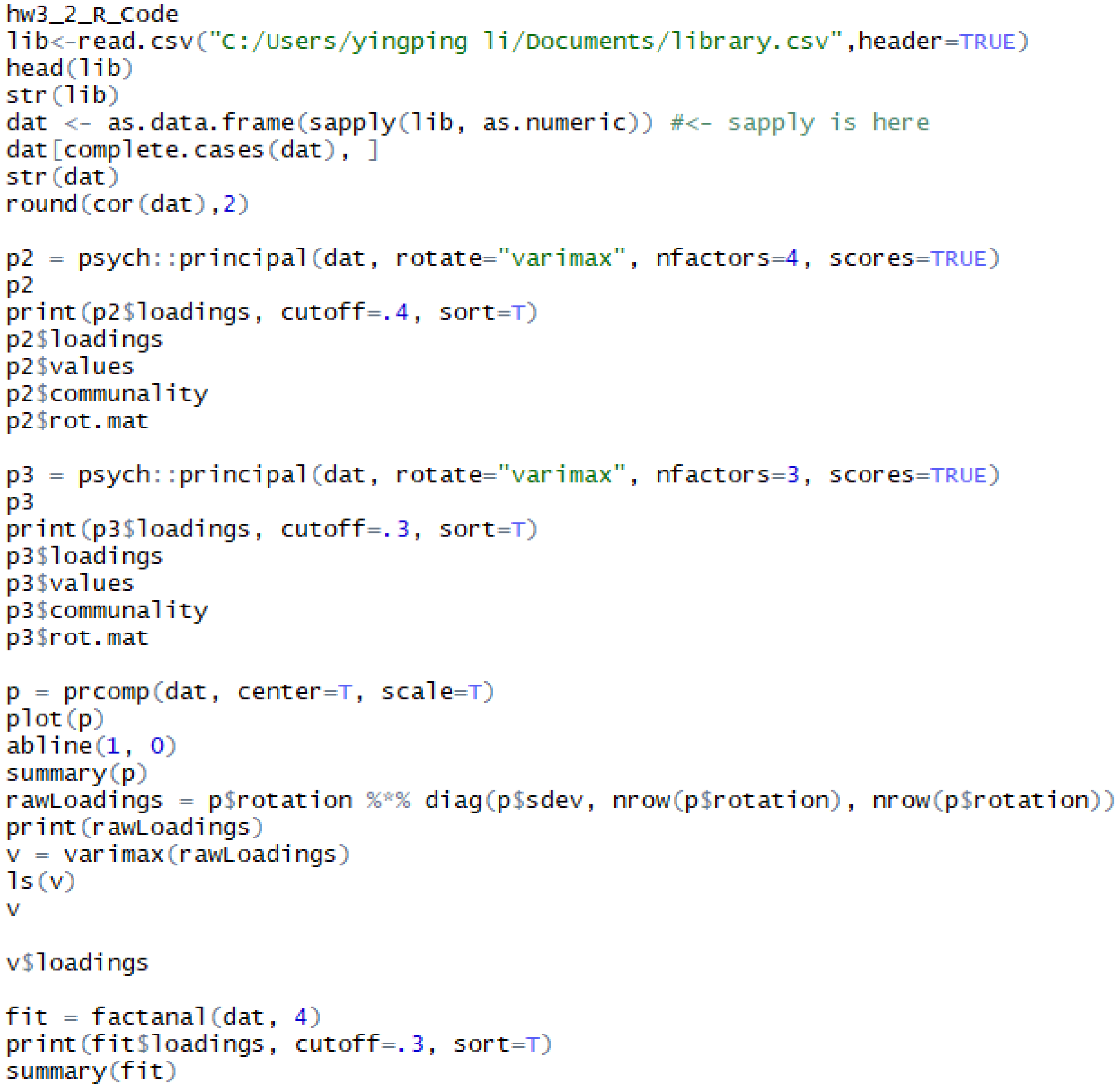
In the future work, I want to find if there are some changes in library resources., visits, and revenues through the analysis of the future library survey data. Also, I want to find which library has the most books per capita.

References

1. Public Library Data Set: https://www.kaggle.com/imls/public-libraries/data
2. “Public Libraries in the United States Survey: Fiscal Year 2014”. Institute of Museum and Library Services. https://www.imls.gov/sites/default/files/publications/documents/plsfy2014.pdf
3. Ian Reid, Ian. “FEATURE: The 2015 Public Library Data Service Statistic Report”. June 22, 2016. http://publiclibrariesonline.org/2016/06/featurethe-2015-public- library-data-service-statistical-report-characteristics-trends/

Appendix

Figure 23: R code for PCA and CFA



```
hw3_2_R_Code
lib<-read.csv("C:/Users/yingping li/Documents/library.csv",header=TRUE)
head(lib)
str(lib)
dat <- as.data.frame(sapply(lib, as.numeric)) #<- sapply is here
dat[complete.cases(dat), ]
str(dat)
round(cor(dat),2)

p2 = psych::principal(dat, rotate="varimax", nfactors=4, scores=TRUE)
p2
print(p2$loadings, cutoff=.4, sort=T)
p2$loadings
p2$values
p2$communality
p2$rot.mat

p3 = psych::principal(dat, rotate="varimax", nfactors=3, scores=TRUE)
p3
print(p3$loadings, cutoff=.3, sort=T)
p3$loadings
p3$values
p3$communality
p3$rot.mat

p = prcomp(dat, center=T, scale=T)
plot(p)
abline(1, 0)
summary(p)
rawLoadings = p$rotation %*% diag(p$sdev, nrow(p$rotation), nrow(p$rotation))
print(rawLoadings)
v = varimax(rawLoadings)
ls(v)
v

v$loadings

fit = factanal(dat, 4)
print(fit$loadings, cutoff=.3, sort=T)
summary(fit)
```