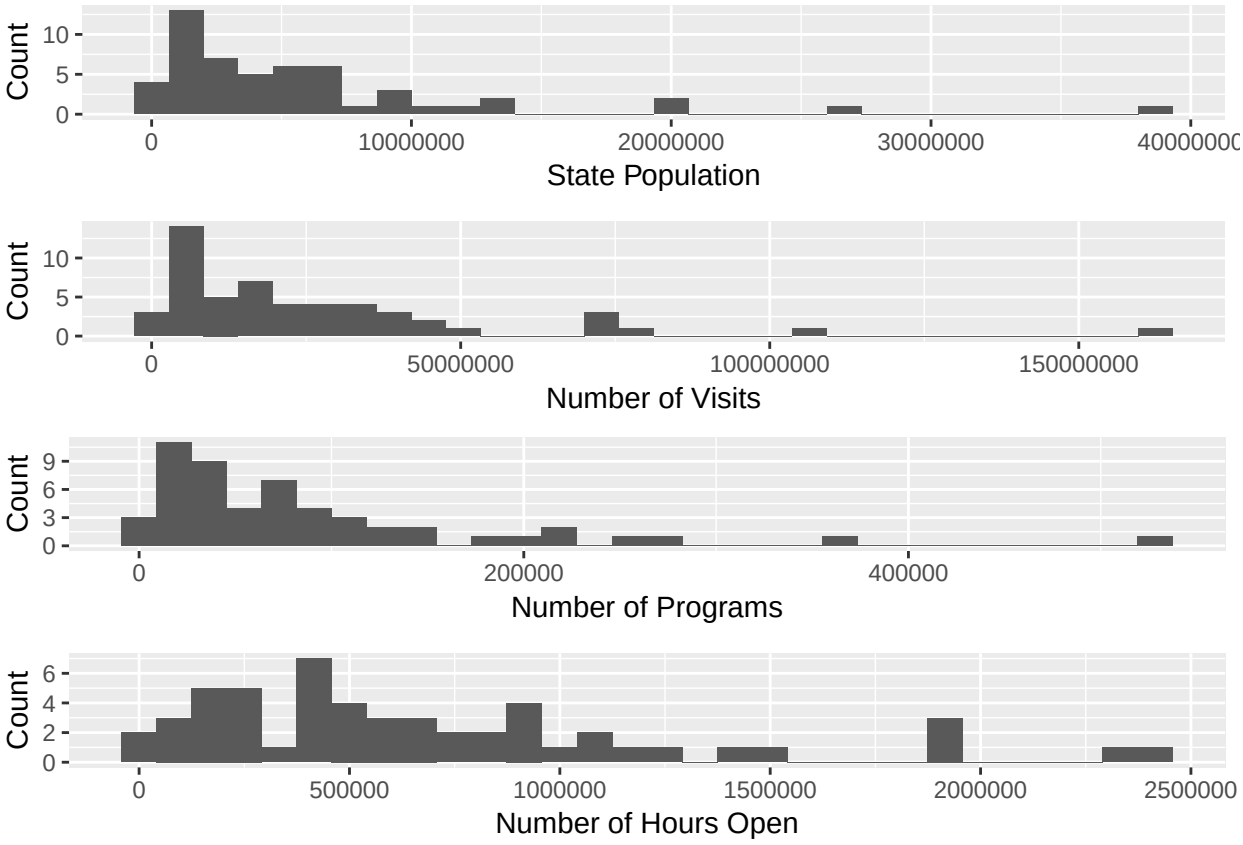
Libraries: What Drives Visits And What Visitors Do by Marie Jordan

==========

The dataset I chose is the Public Library State Summary/State Characteristics data file from FY2015, provided by the Institute of Museum and Library Services. The information comes from all public libraries in the United States including Washington DC, Guam and American Samoa. The data was accompanied by an extensive document detailing collection techniques and data manipulation, as well as defining the terms used in the table.

It is worth noting that this dataset is collected data from the entire population of public libraries in the US. It includes imputations for non- responding libraries to minimize the effects of nonresponse. Other relevant data is that there are different reporting periods from state to state (although all provided data for at least 12 months) and that there is different adherence to the definitions of collected data from state to state.

Univariate Plots Section

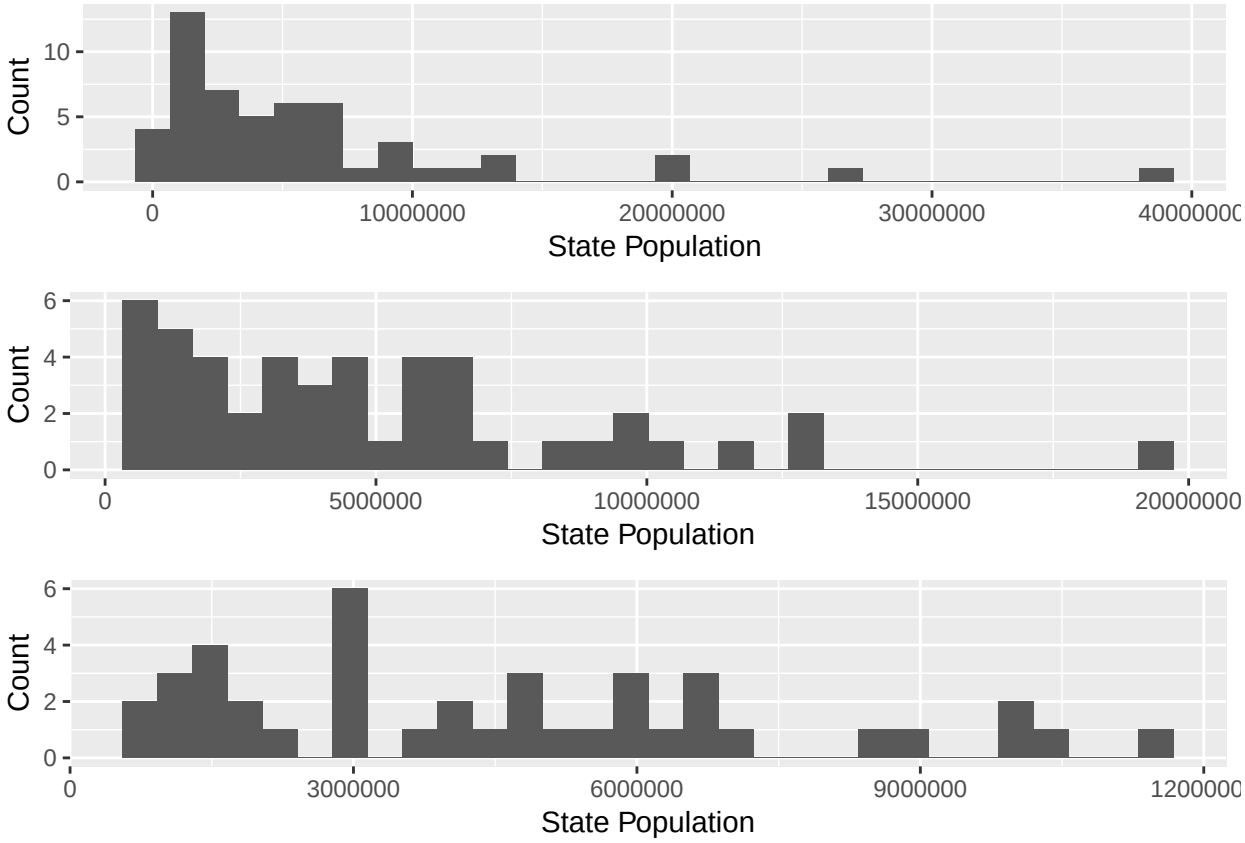


This is a quick-and-dirty look for any interesting patterns in the variables I'm examining that are related to visits to the libraries.

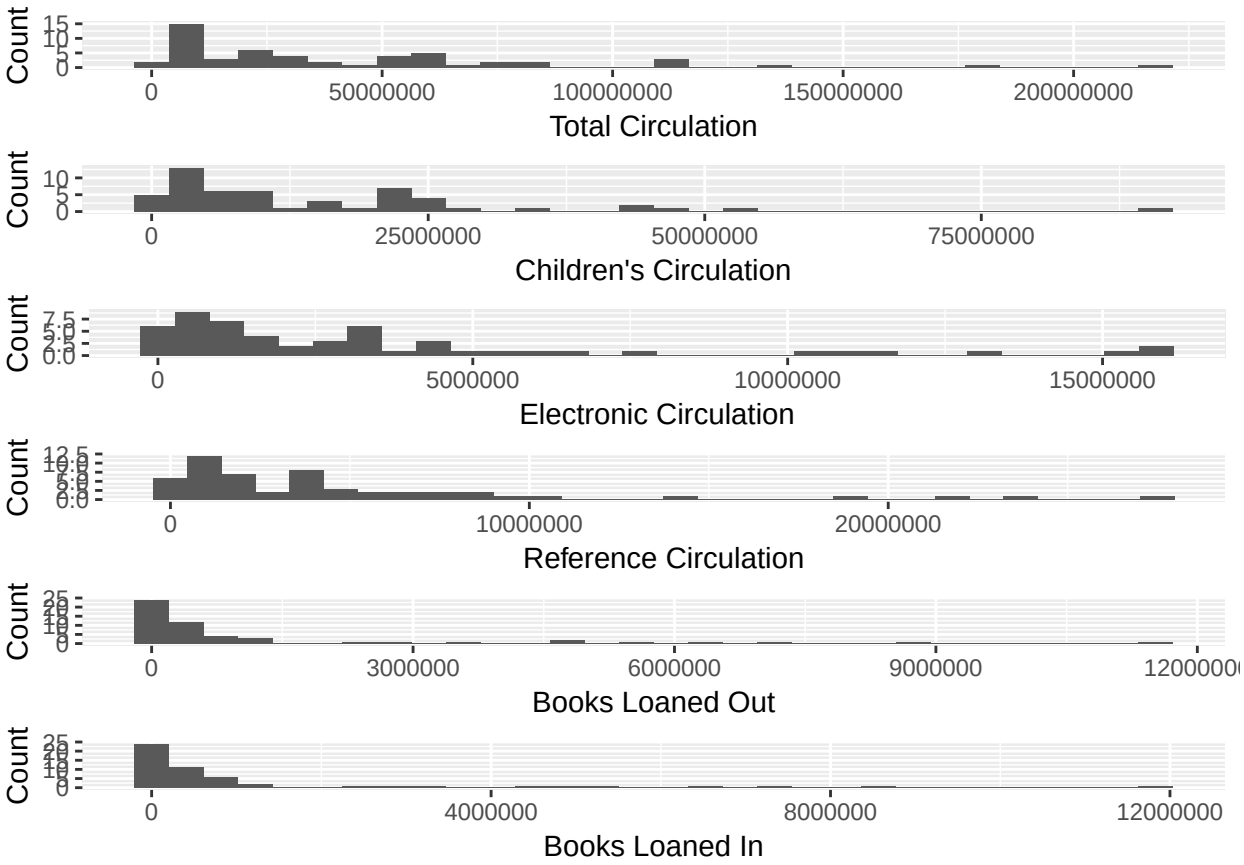
```
##      STABR       POPU_ST             VISITS              TOTPRO
##  AK     : 1   Min.   :   60863   Min.   :    63563   Min.   :   384
##  AL     : 1   1st Qu.: 1419561   1st Qu.:  6719983   1st Qu.: 27283
##  AR     : 1   Median : 4013845   Median : 18107047   Median : 58459
##  AS     : 1   Mean   : 6007530   Mean   : 26350902   Mean   : 89501
```

```
##  AZ     : 1   3rd Qu.: 6758251   3rd Qu.: 34932970   3rd Qu.:105285
##  CA     : 1   Max.   :38714725   Max.   :162526811   Max.   :528590
##  (Other):47
##     HRS_OPEN
##  Min.   :   3523
##  1st Qu.: 271535
##  Median : 510862
##  Mean   : 699438
##  3rd Qu.: 921095
##  Max.   :2418181
##
```

Summary table for visit-related variables.



These plots show state population numbers with outliers removed. (In this case, because the data is arranged about the same across all variables, population is acting as a test case.) The topmost plot has all 53 datapoints, the middle plot removes the top and bottom 5%, and the bottom plot removes the top and bottom 10%.



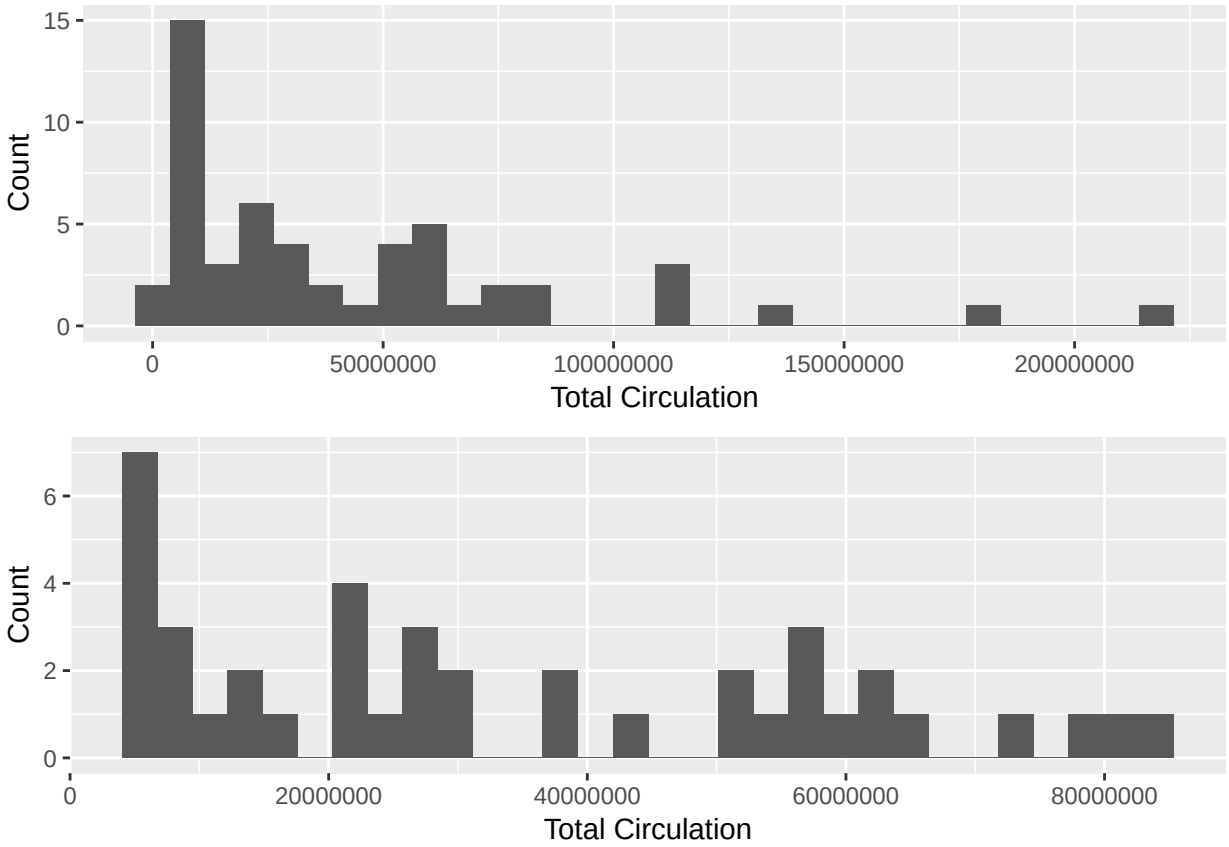
To cover visitor interactions at the library, I created a dataframe with state names, population numbers, and variables related to visitor actions in the libraries.

These tables are a quick-and-dirty look for interesting patterns. As expected, circulation of electronic materials and kid's materials is less than the total number. Loans to and from seem to echo each other in very interesting ways – states seem to have the same numbers for loans to and loans from, for the most part.

```
##      STABR       POPU_ST             TOTCIR             KIDCIRCL
##  AK     : 1   Min.   :   60863   Min.   :    22280   Min.   :    9182
##  AL     : 1   1st Qu.: 1419561   1st Qu.:  7673400   1st Qu.: 2729594
##  AR     : 1   Median : 4013845   Median : 26302661   Median : 9634627
##  AS     : 1   Mean   : 6007530   Mean   : 42848660   Mean   :15132834
##  AZ     : 1   3rd Qu.: 6758251   3rd Qu.: 58275229   3rd Qu.:21903733
##  CA     : 1   Max.   :38714725   Max.   :217860542   Max.   :90787494
##  (Other):47
##     ELMATCIR           REFERENC            LOANTO
##  Min.   :      -1   Min.   :    1154   Min.   :       0
##  1st Qu.:  642244   1st Qu.:  790724   1st Qu.:   52914
##  Median : 2259776   Median : 2804740   Median :  301352
##  Mean   : 3626635   Mean   : 4833300   Mean   : 1324420
##  3rd Qu.: 4433382   3rd Qu.: 5753809   3rd Qu.:  993074
##  Max.   :15852731   Max.   :27492044   Max.   :11523847
##
##      LOANFM
##  Min.   :       0
##  1st Qu.:   45897
##  Median :  323486
```

```
##  Mean   : 1329872
##  3rd Qu.:  960202
##  Max.   :11833557
##
```

Summary of the table of actions patrons took at the library.



As with the physical data subset, there are huge differences between the min and max, and the 3rd quartile and max. The plot shows the dataset for total circulation with the top and bottom 10% of states removed, to check for patterns without outliers.

Univariate Analysis

What is the structure of your dataset?

My dataset is large, with 53 observations of 124 different variables. It's arranged by state, with each row containing the state name, and then all of the related information.

For this analysis, I'm concentrating on number of visits, and what visitors do once they're in the libraries. The dataset is notable for the huge range of data it contains – I could as easily have looked at funding streams, wages and benefits, expenses, population served, or systems of main and branch libraries and the staffing patterns found therein.

What is/are the main feature(s) of interest in your dataset?

My data explorations are going to tackle two primary questions – what drives number of visits (focusing on number of programs and hours open in a year), and what they do once they're there (borrow books or electronic resources , use the library for their children or themselves, file reference requests, etc.) I am examining these numbers broken down by state.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

The variables I've chosen to look at directly relate to the two major interest features I'm examining. Although there are likely other features in the dataset that have influence, I've chosen to primarily look at hours open and programs put on to influence visit numbers, and circulation, reference requests and loans as aspects of what library users actually do.

Did you create any new variables from existing variables in the dataset?

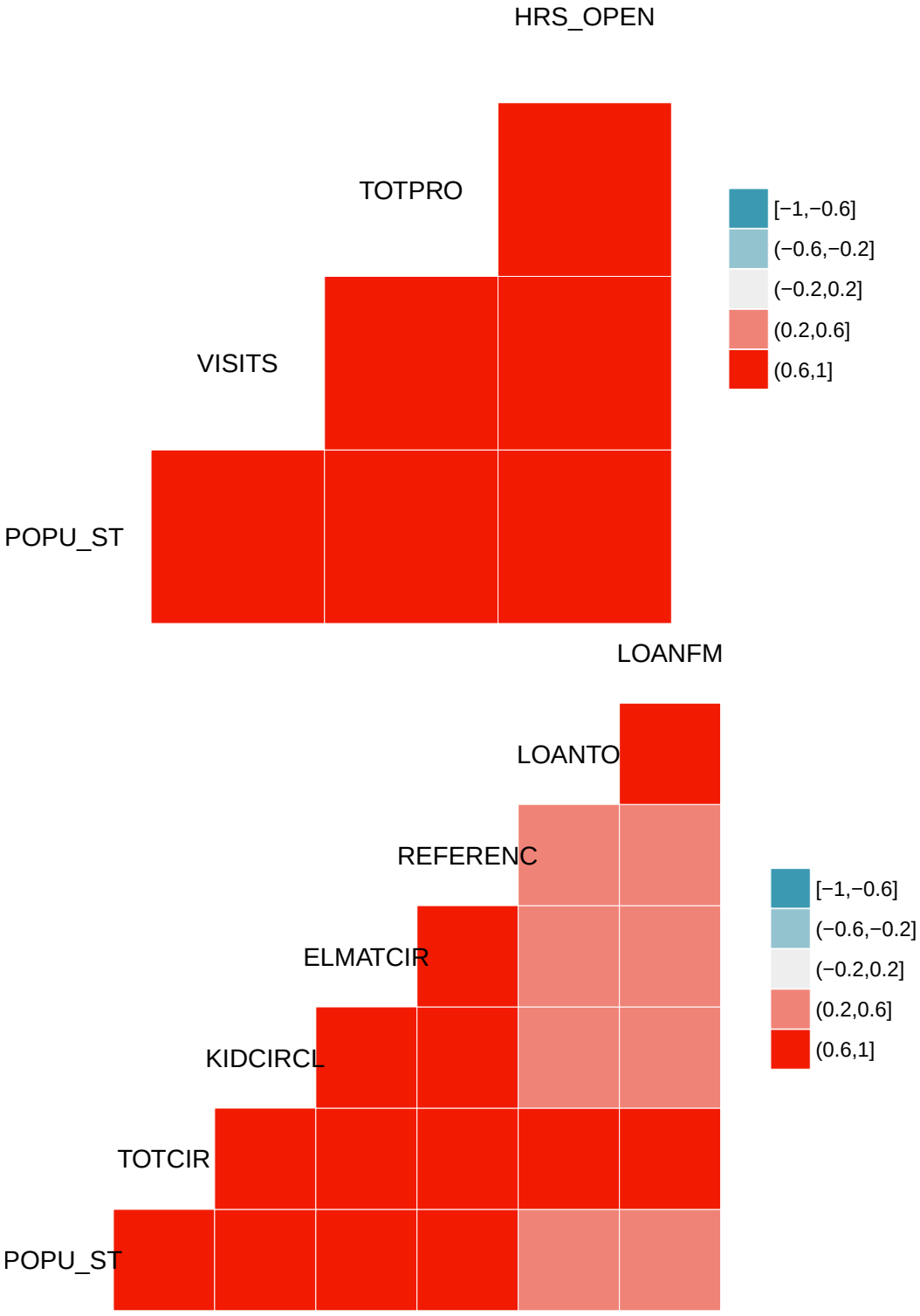
I did not create any new variables at this stage, nor did I transform or tidy the data, beyond creating two new dataframes. For my visit numbers information, I created a new dataframe called physical which breaks out the variables I'm interested in. This makes working with the data easier, especially for things like summaries. I performed a similar action for visitor activities, creating a new dataframe called action, for the same reasons I created the physical dataframe.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

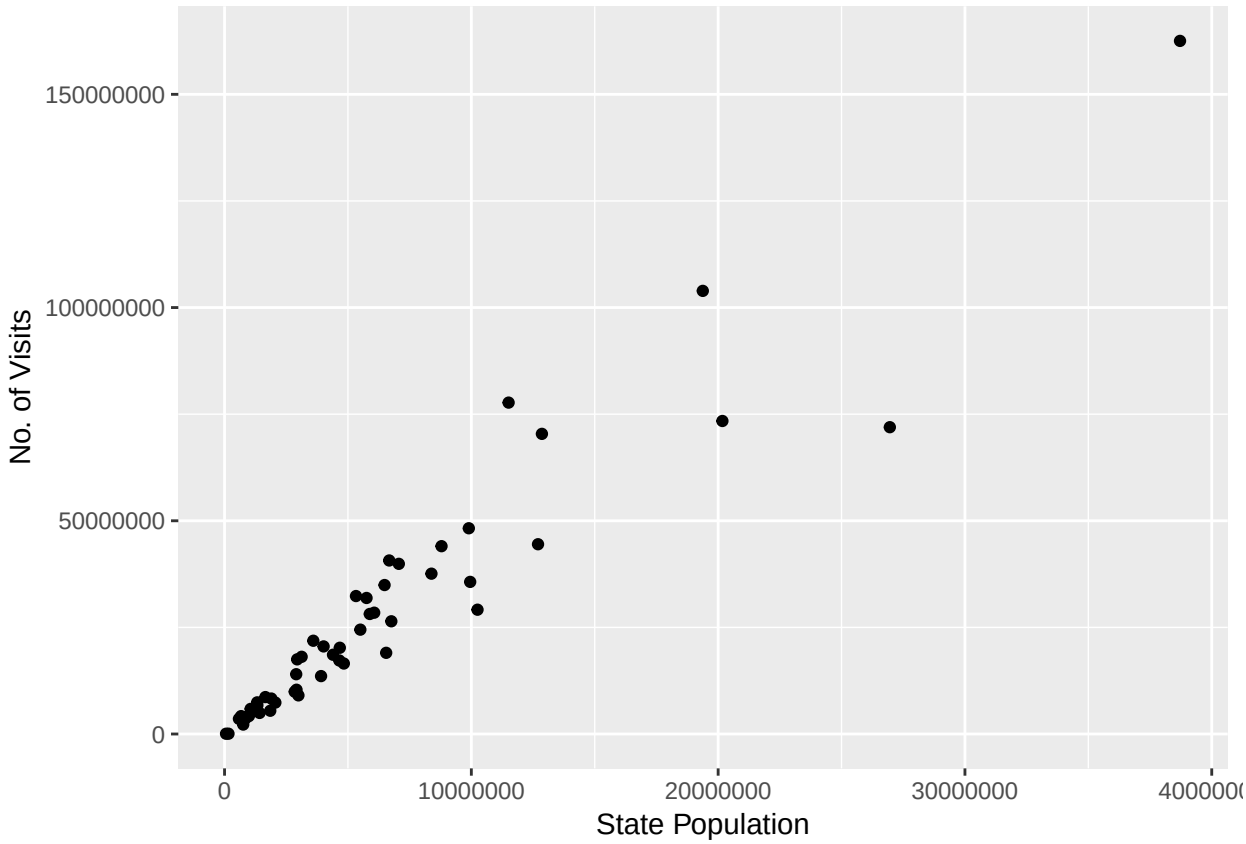
With my analyses, I did not find any particularly unusual distributions, only that there was a huge difference between minimum and maximum numbers for all variables studied, as would be expected when working with state-level data. The majority of the data for all variables was distributed roughly evenly clustered at the lower end of the scale, with only a few larger outliers, generally in high-population states.

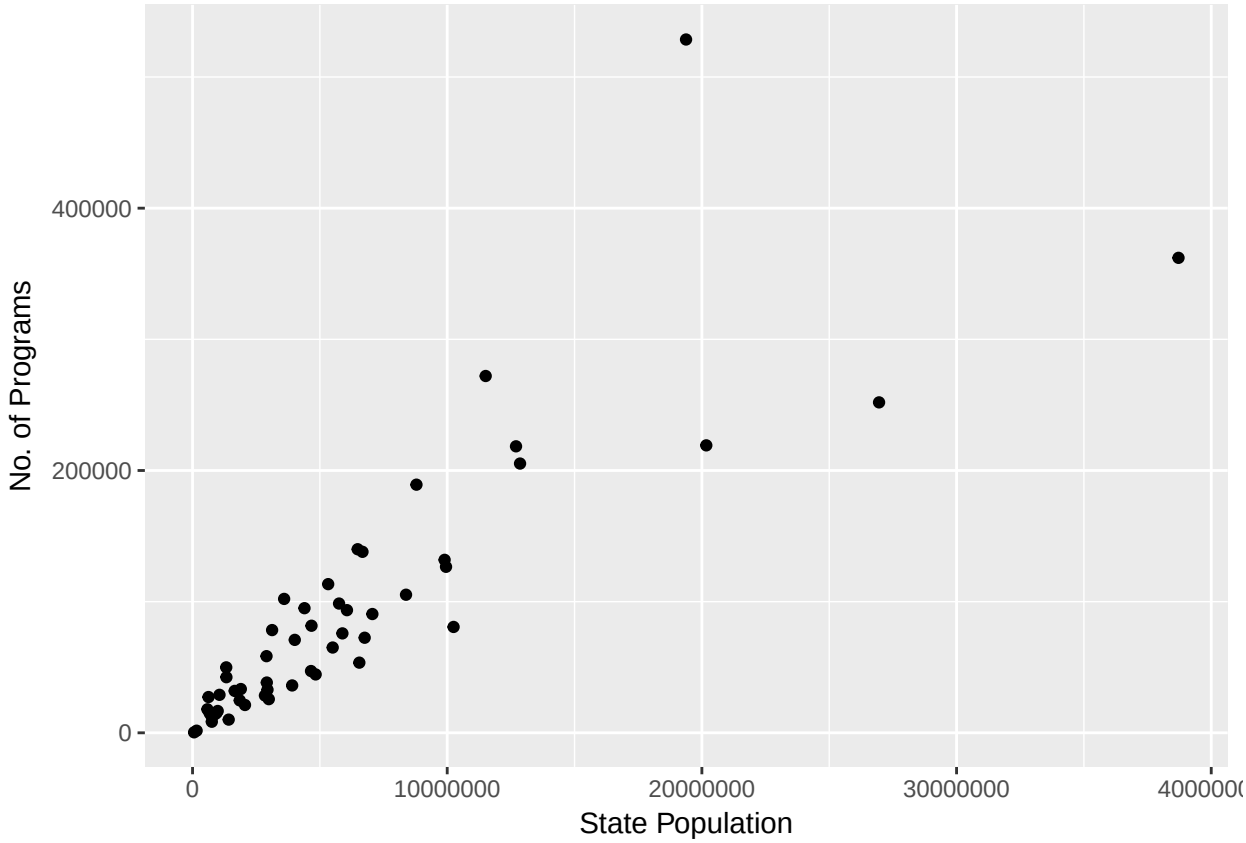
Removing the outliers did not reveal any new patterns or throw up anything of particular interest.

Bivariate Plots Section

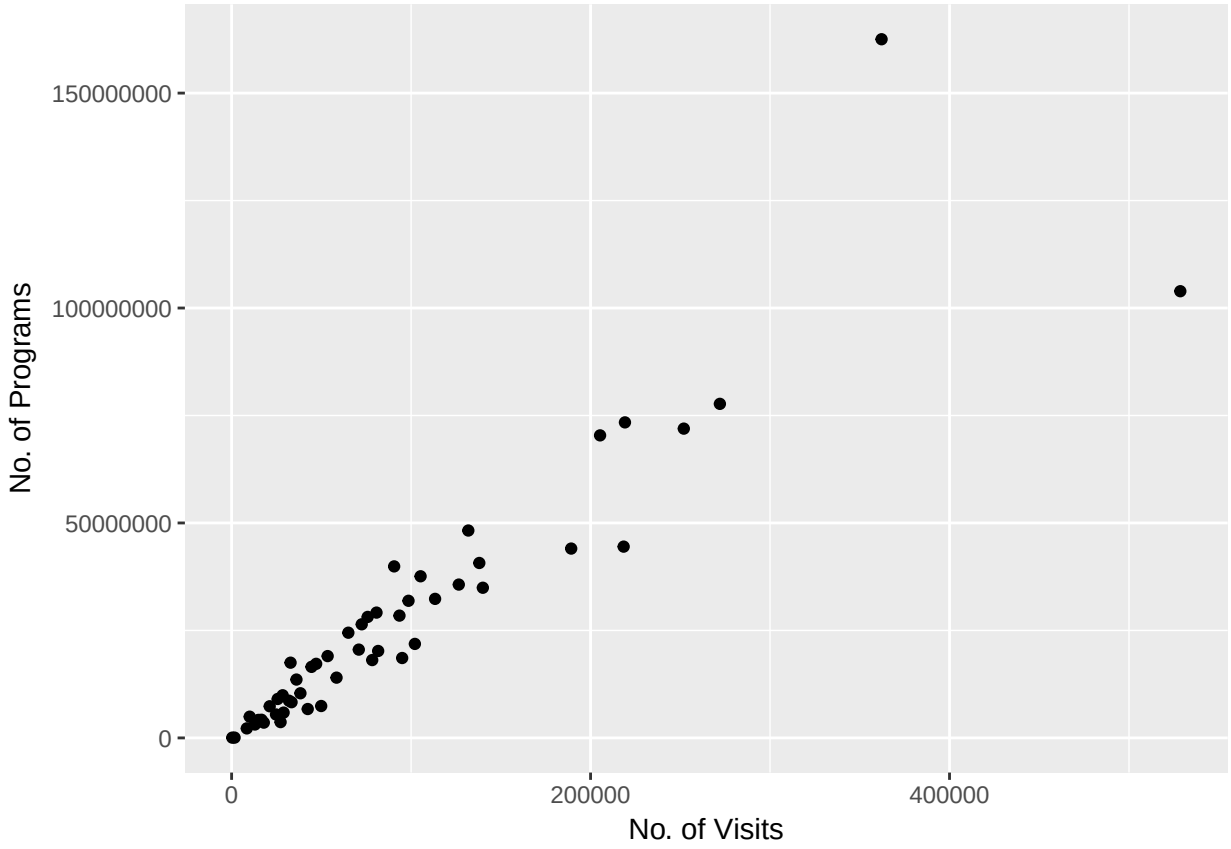


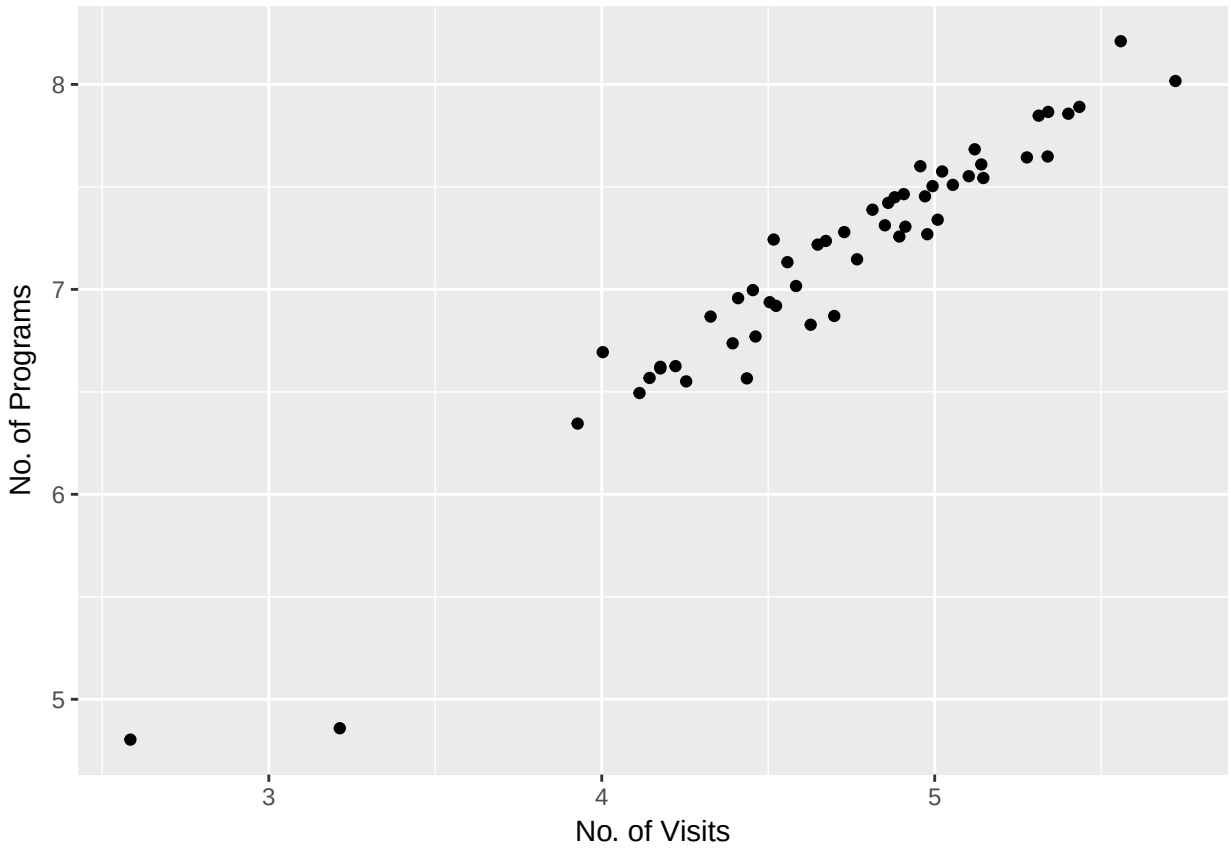
Correlation plots show potential correlations between all variables examined.


150000000
100000000
50000000
0
No. of Visits
0
10000000
20000000
30000000
State Population

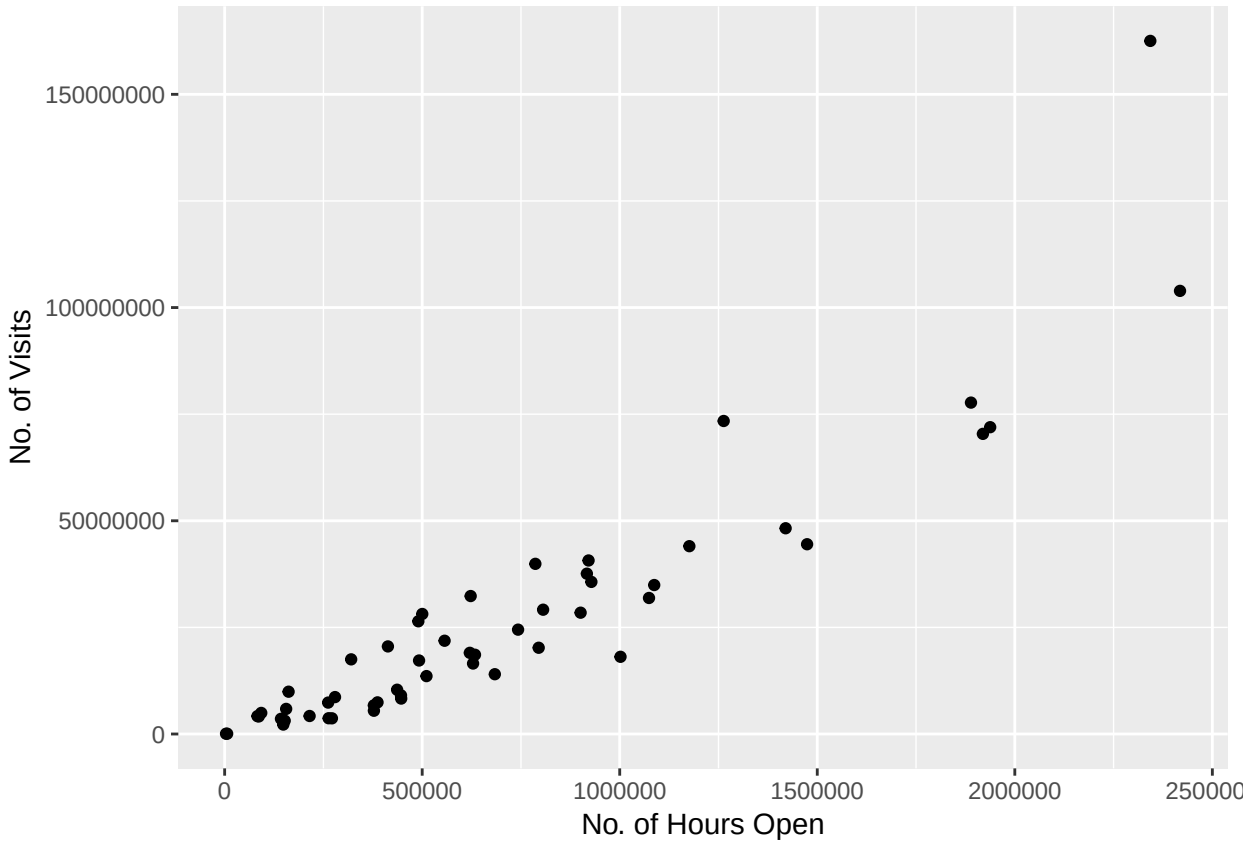

400000
200000
0
No. of Programs
0
10000000
20000000
30000000
State Population

These plots show the relationship between state populations and the number of visits and number of programs in the submitted timeframe.





These plots examine the relationship between the number of programs and the number of visits in a state. I did a version with the log10 of both variables, despite the obvious linear regression in the first graph, simply because I'm still learning how log10 works and I was curious to see what effect, if any, it would have on the data.



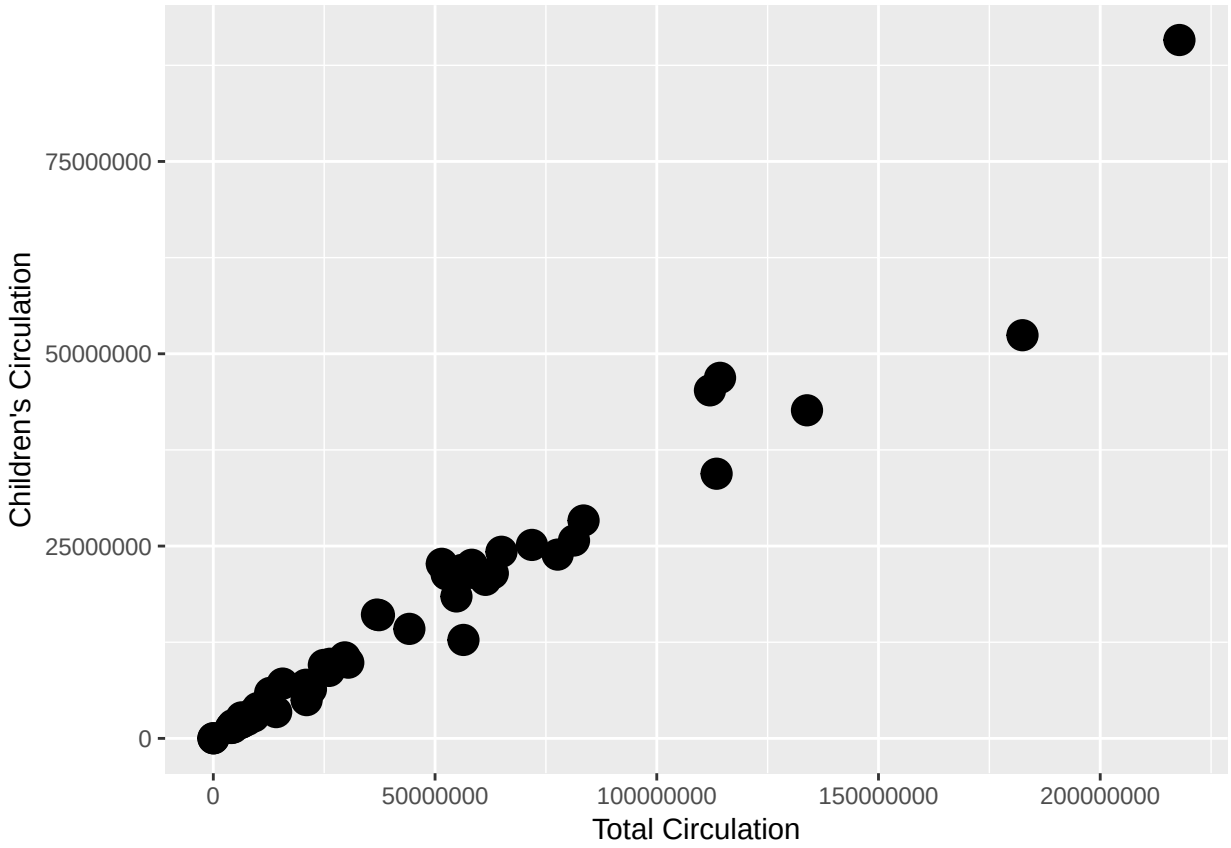
As above, this plot examines the relationship between hours open and number of visits.

```
##
## Call:
## lm(formula = VISITS ~ TOTPRO, data = physical)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -43342933  -3087886   -889828   2562689  61106094
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 1706377.31 2221905.81   0.768               0.446
## TOTPRO          275.35      16.77  16.415 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11920000 on 51 degrees of freedom
## Multiple R-squared:  0.8408, Adjusted R-squared:  0.8377
## F-statistic: 269.5 on 1 and 51 DF,  p-value: < 0.00000000000000022
```

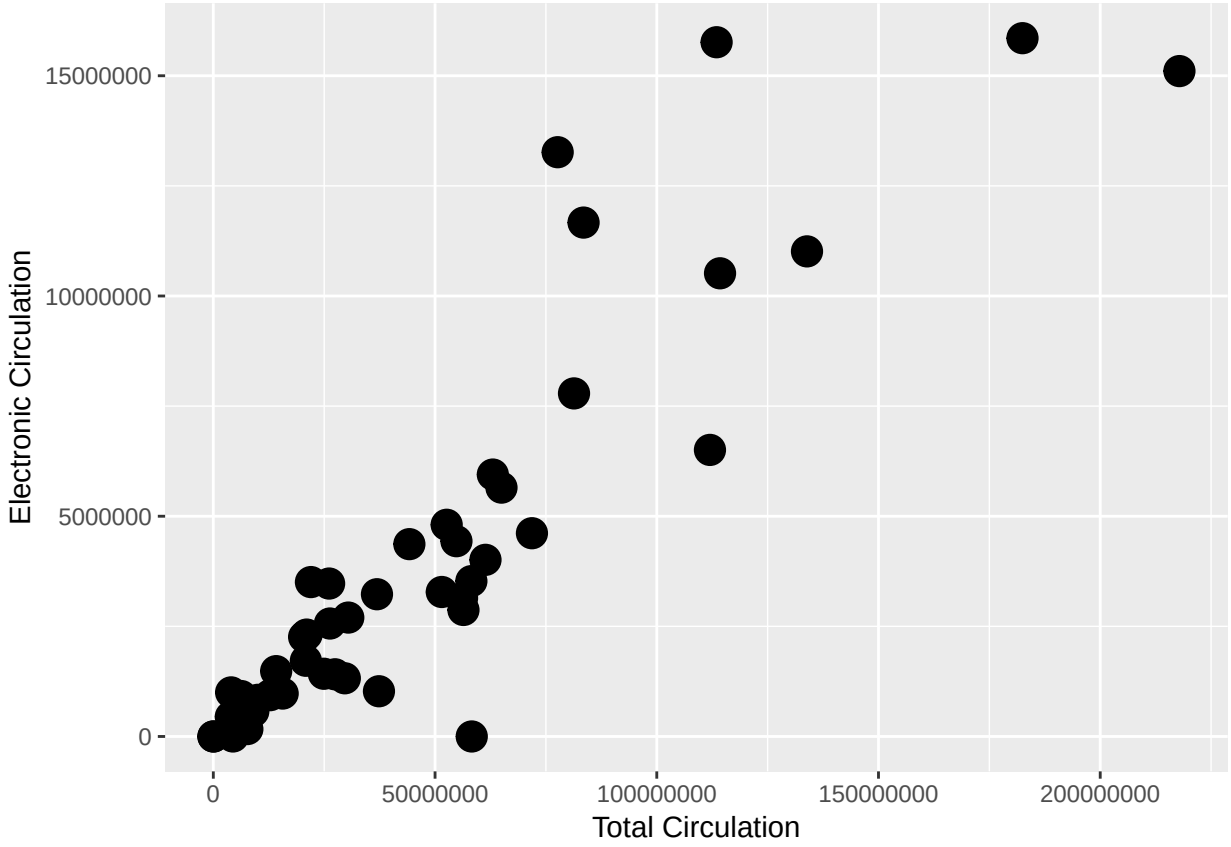
```
##
## Call:
## lm(formula = VISITS ~ HRS_OPEN, data = physical)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -22304572  -5530724  -2161684   5938737  59781593
```

```
##
## Coefficients:
##                 Estimate   Std. Error t value            Pr(>|t|)
## (Intercept) -6159200.357  2519572.539  -2.445               0.018 *
## HRS_OPEN          46.480        2.772  16.767 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11710000 on 51 degrees of freedom
## Multiple R-squared:  0.8464, Adjusted R-squared:  0.8434
## F-statistic: 281.1 on 1 and 51 DF,  p-value: < 0.00000000000000022
```

Linear regression modeling confirms a significant relationship between both number of programs and hours open and number of visits.



This plot compares how total circulation numbers predict kid circulation numbers. There's a clear linear relationship between the two, with kid circulation going up as total circulation does.



Similar to the plot above, this one looks at how total circulation relates to electronic circulation. The relationship is less obvious this time.

```
##
## Call:
## lm(formula = KIDCIRCL ~ TOTCIR, data = action)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -12730301   -903333    124024    830083  12970081
##
## Coefficients:
##                   Estimate    Std. Error t value            Pr(>|t|)
## (Intercept) -214410.82792  653922.09045  -0.328               0.744
## TOTCIR            0.35817       0.01044  34.292 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3471000 on 51 degrees of freedom
## Multiple R-squared:  0.9584, Adjusted R-squared:  0.9576
## F-statistic:  1176 on 1 and 51 DF,  p-value: < 0.00000000000000022

##
## Call:
## lm(formula = ELMATCIR ~ TOTCIR, data = action)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -4928152  -490543   -53921   375655  6701114
##
## Coefficients:
##                  Estimate    Std. Error t value            Pr(>|t|)
## (Intercept)  11560.552302 351294.527944   0.033               0.974
## TOTCIR           0.084368      0.005611  15.036 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1865000 on 51 degrees of freedom
## Multiple R-squared:  0.8159, Adjusted R-squared:  0.8123
## F-statistic: 226.1 on 1 and 51 DF,  p-value: < 0.00000000000000022
```

Linear regression models. The first one examines the effect total circulation numbers have on circulation of children's materials, and the second one looks at the relationship total circulation has on the circulation of electronic materials.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

The features I examined had fairly straightforward relationships. They generally had very strong linear relationships, with the y axis variable rising as the x axis variable does so. The most interesting relationships were the ones that didn't follow a strong linear trend! Circulation of electronic materials graphed against total circulation results in a noiser scatterplot than almost any other I looked at. This would be an interesting avenue for investigation, especially by someone interested in the digital divide, or in how people in different states use their library system.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

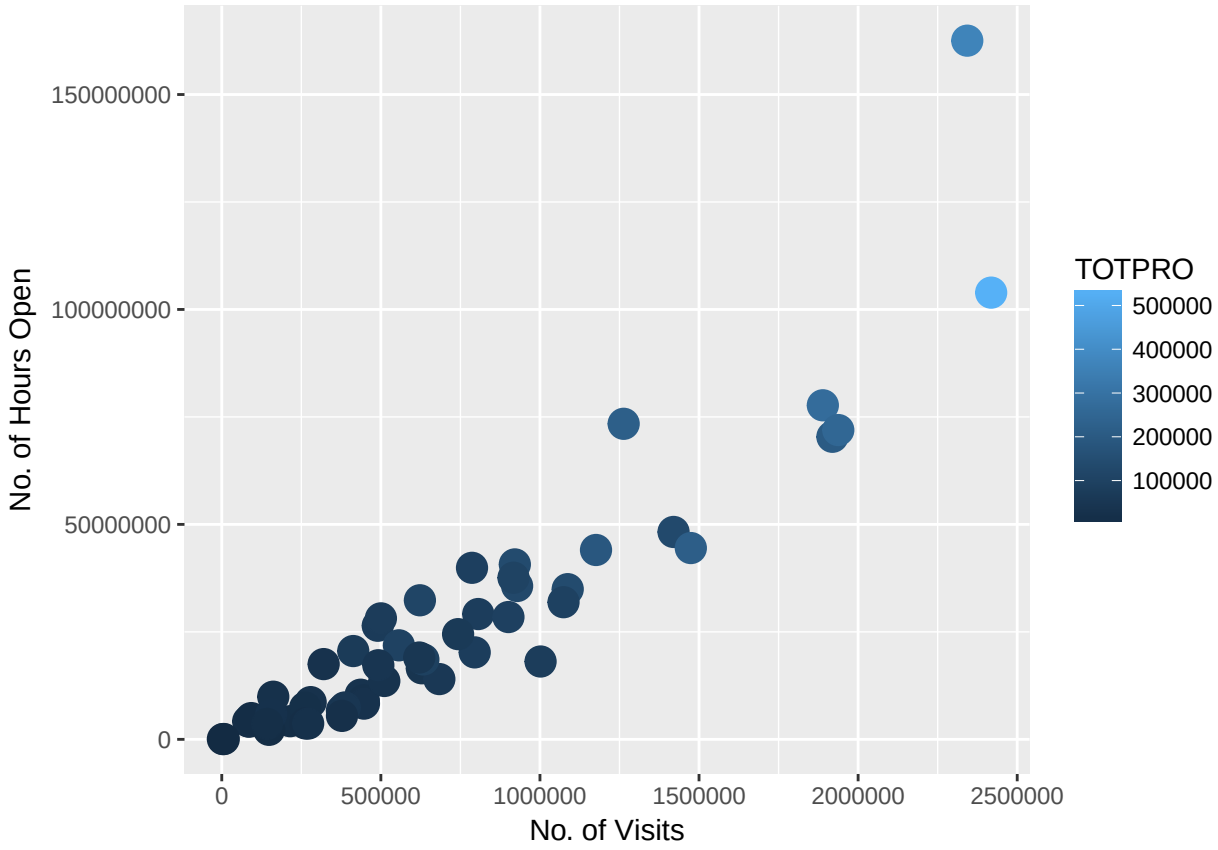
There is a strong linear regression relationship between the number of visits and both the hours open and number of programs. Although the log10 transformation didn't reveal anything new, I have a clearer idea of what it can be useful for. In general, as both number of hours open and number of programs offered increases, visit numbers also increase. Conducting a linear regression shows that both variables are significant, with R-squared values near 1.

As with visit data, the circulation data revealed that the tested variables are significant.

What was the strongest relationship you found?

Most of the relationships I studied were quite strong, but the number of hours open influencing the number of visits was particularly powerful.

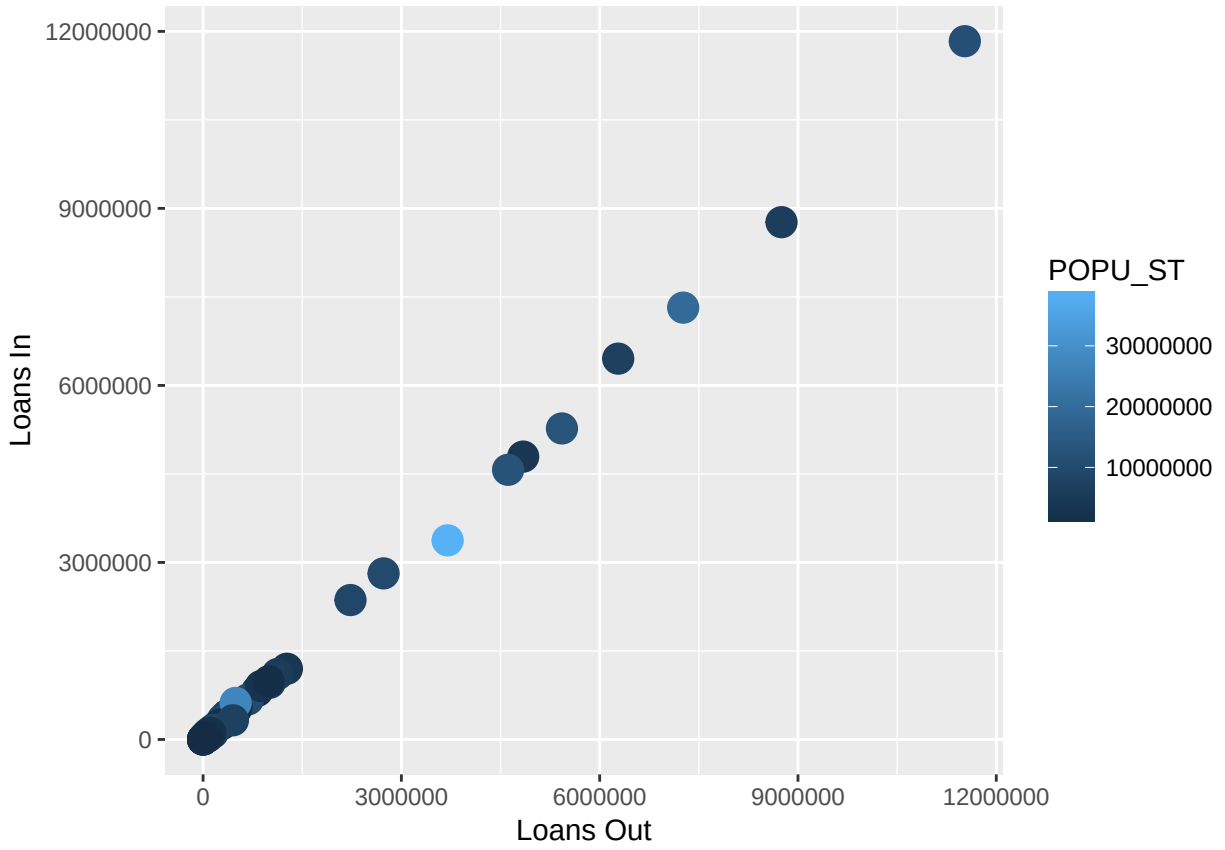
Multivariate Plots Section



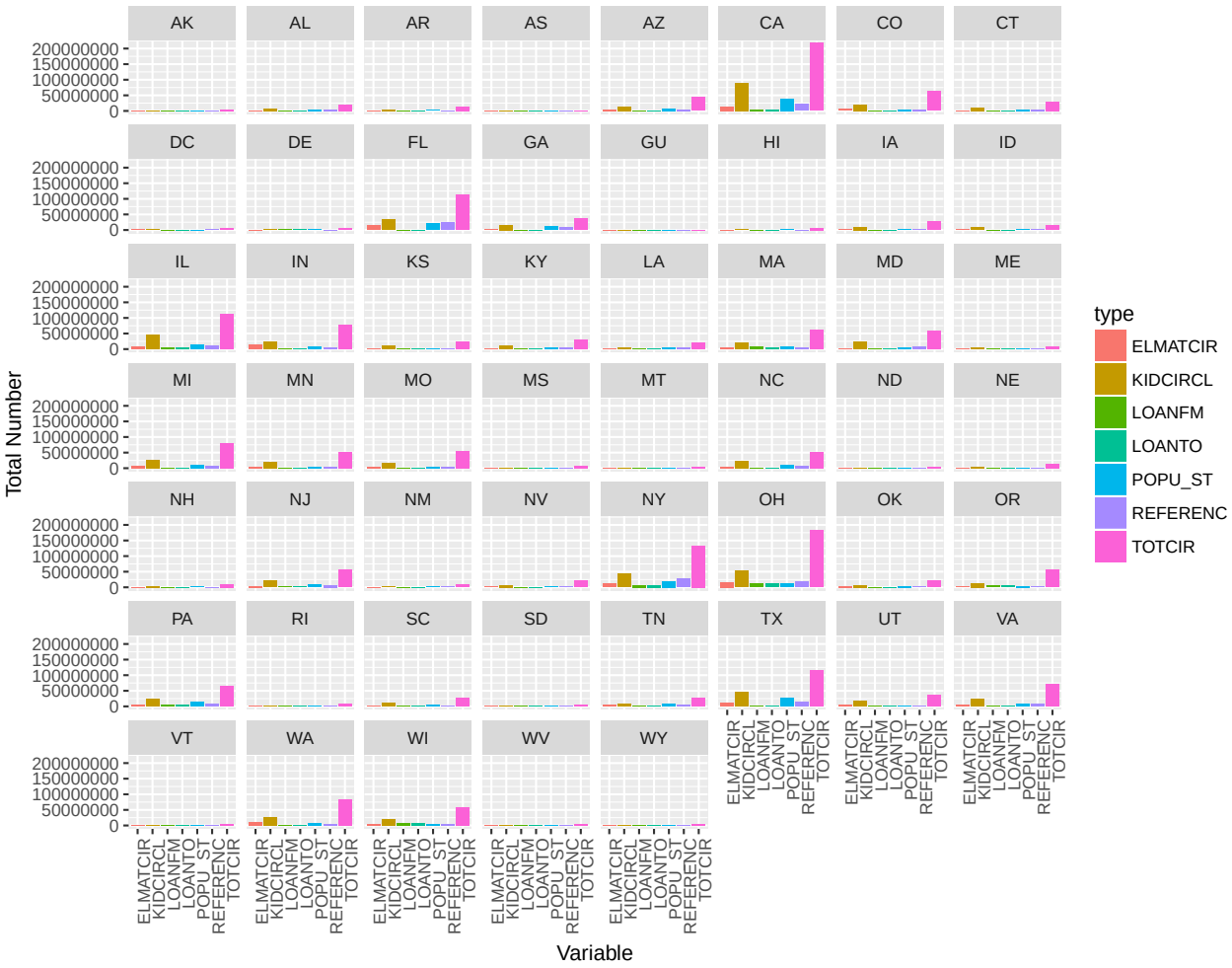
This plot shows the relationship between hours open and visits, colored by number of programs.

```
##
## Call:
## lm(formula = VISITS ~ TOTPRO + HRS_OPEN, data = physical)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -25282869  -4012517   -445925   3301084  57818949
##
## Coefficients:
##                Estimate   Std. Error t value Pr(>|t|)
## (Intercept) -3394485.159  2472177.060  -1.373  0.17586
## TOTPRO           135.753       42.304   3.209  0.00233 **
## HRS_OPEN          25.156        7.117   3.535  0.00089 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10770000 on 50 degrees of freedom
## Multiple R-squared:  0.8727, Adjusted R-squared:  0.8676
## F-statistic: 171.3 on 2 and 50 DF,  p-value: < 0.00000000000000022
```

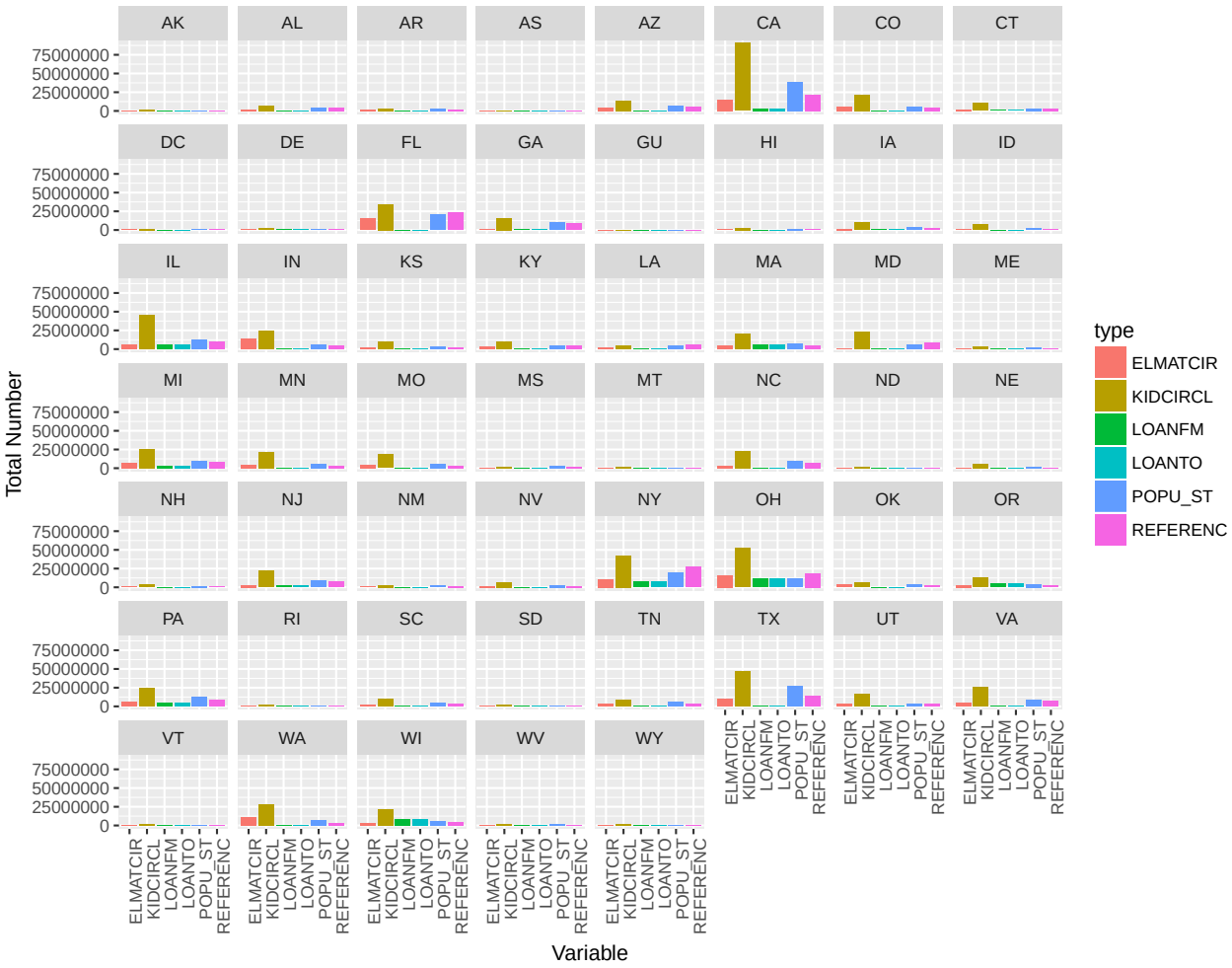
Linear regression analysis for programs and hours open predicting visits.



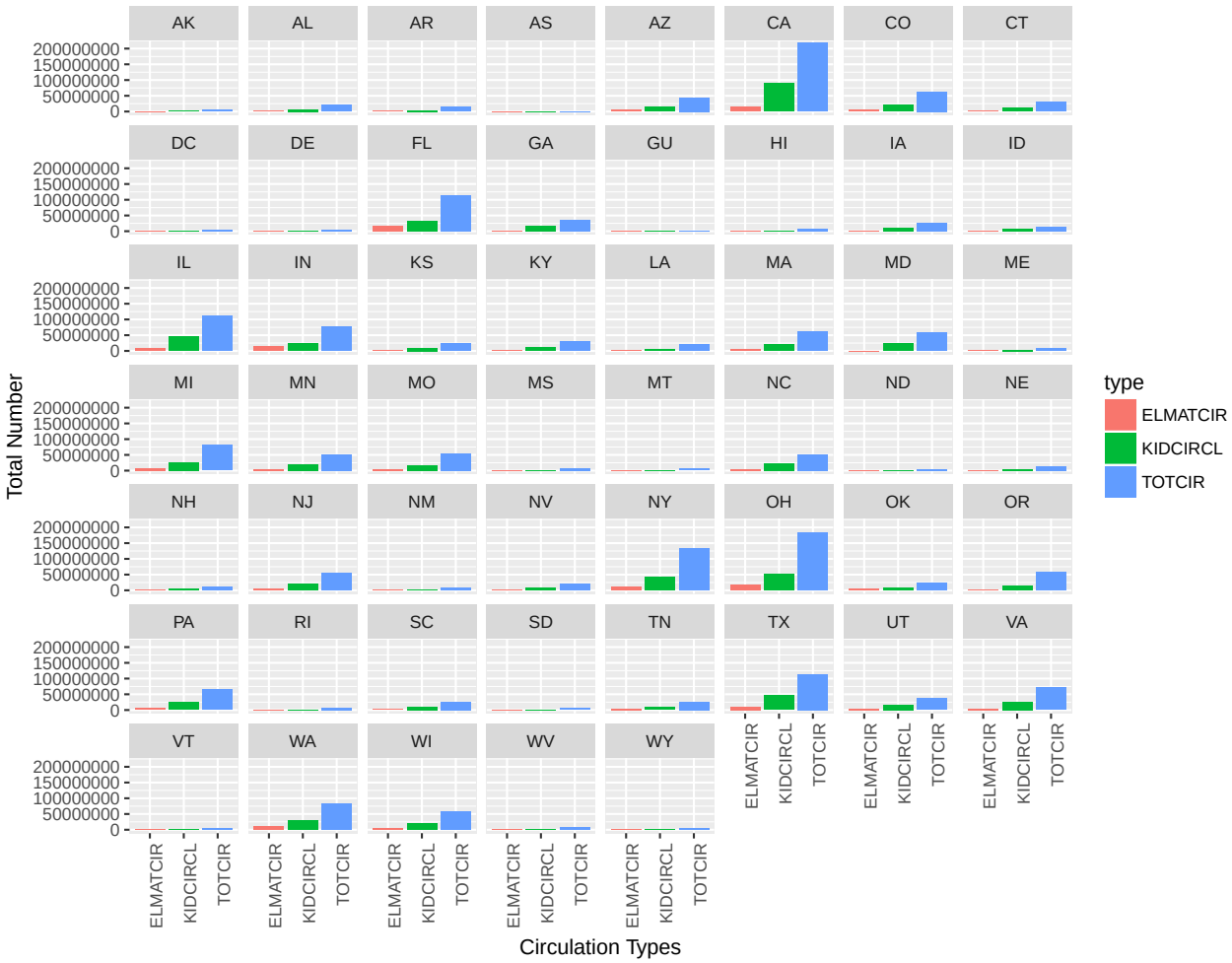
Comparing loan to and loan from, colored by population. The loans numbers are almost completely identical, creating a linear regression probably close to 1. They are not, however, influenced by population size.



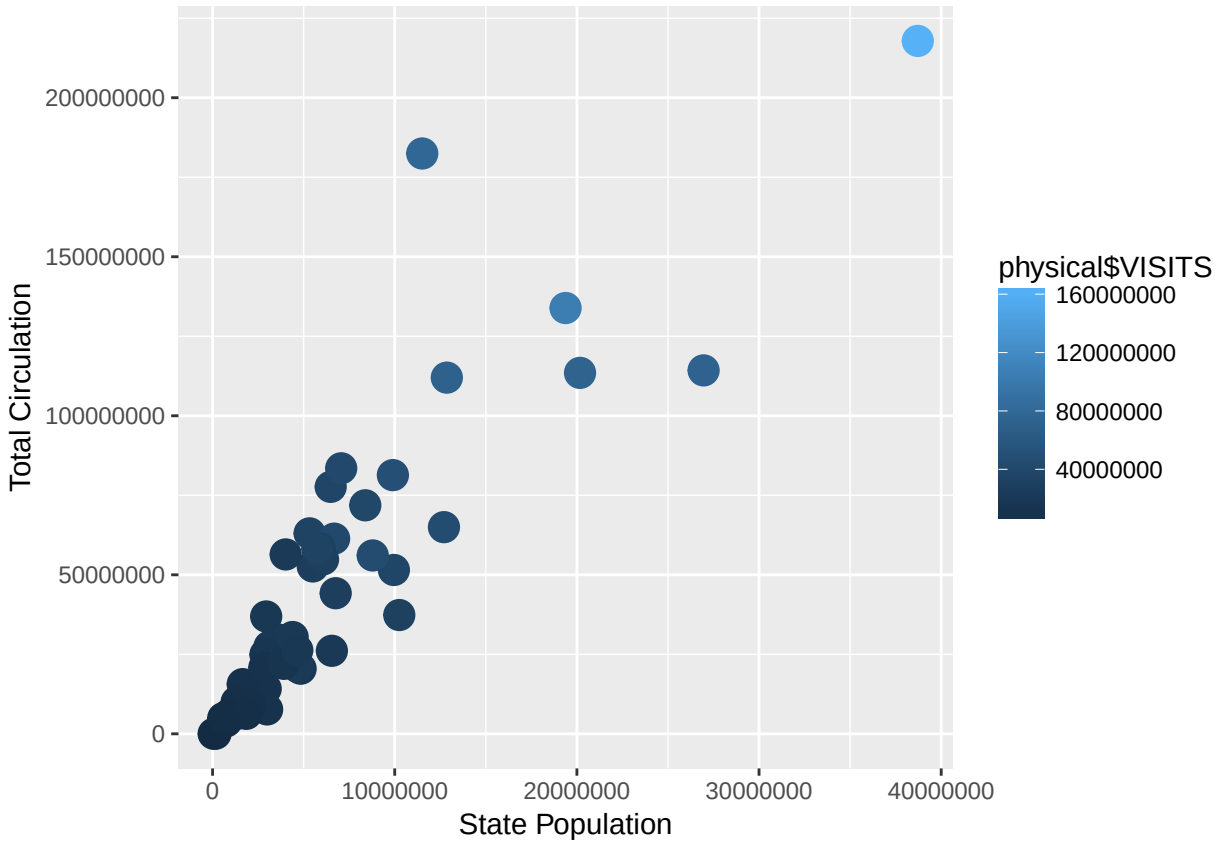
A series of barplots, broken down by state, to compare all variables in the action table.



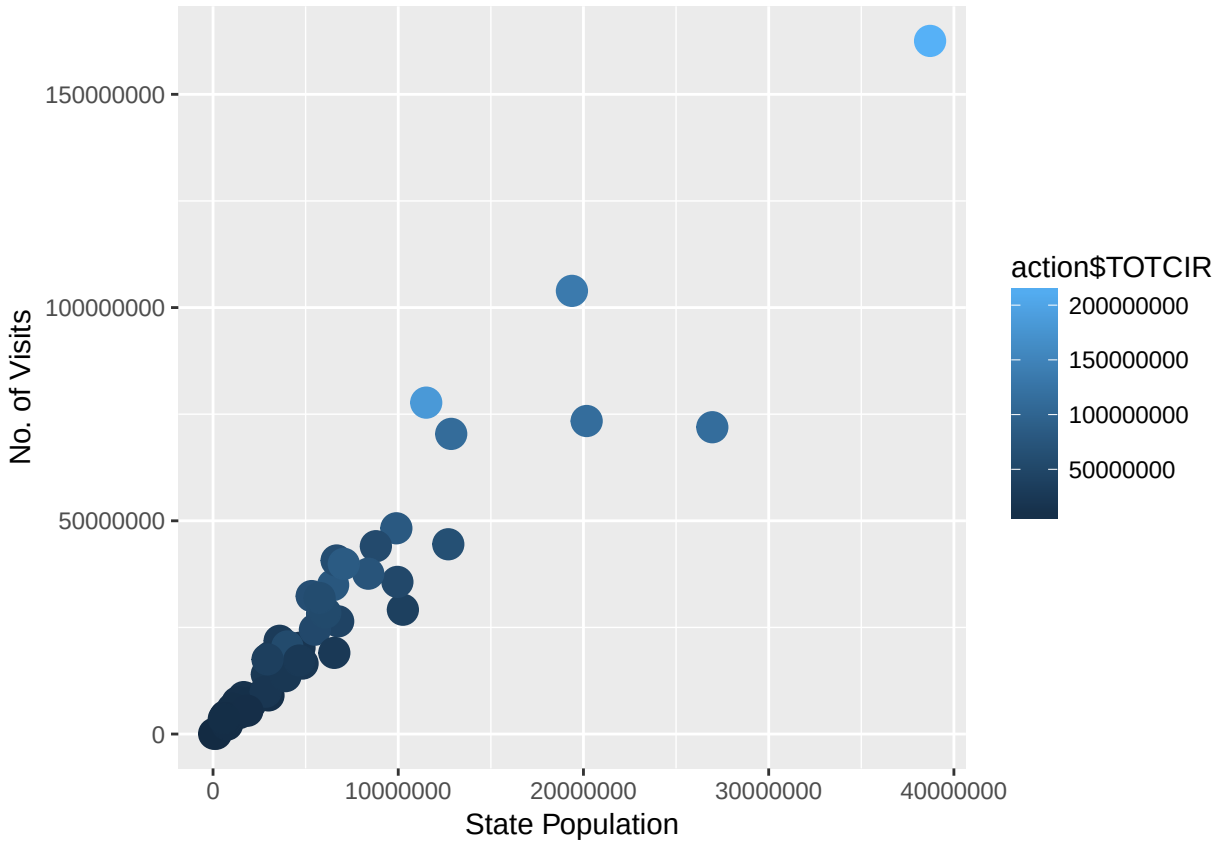
This is the same series of barplots as above, although with total circulation removed to better show the relative volumes of the other variables.



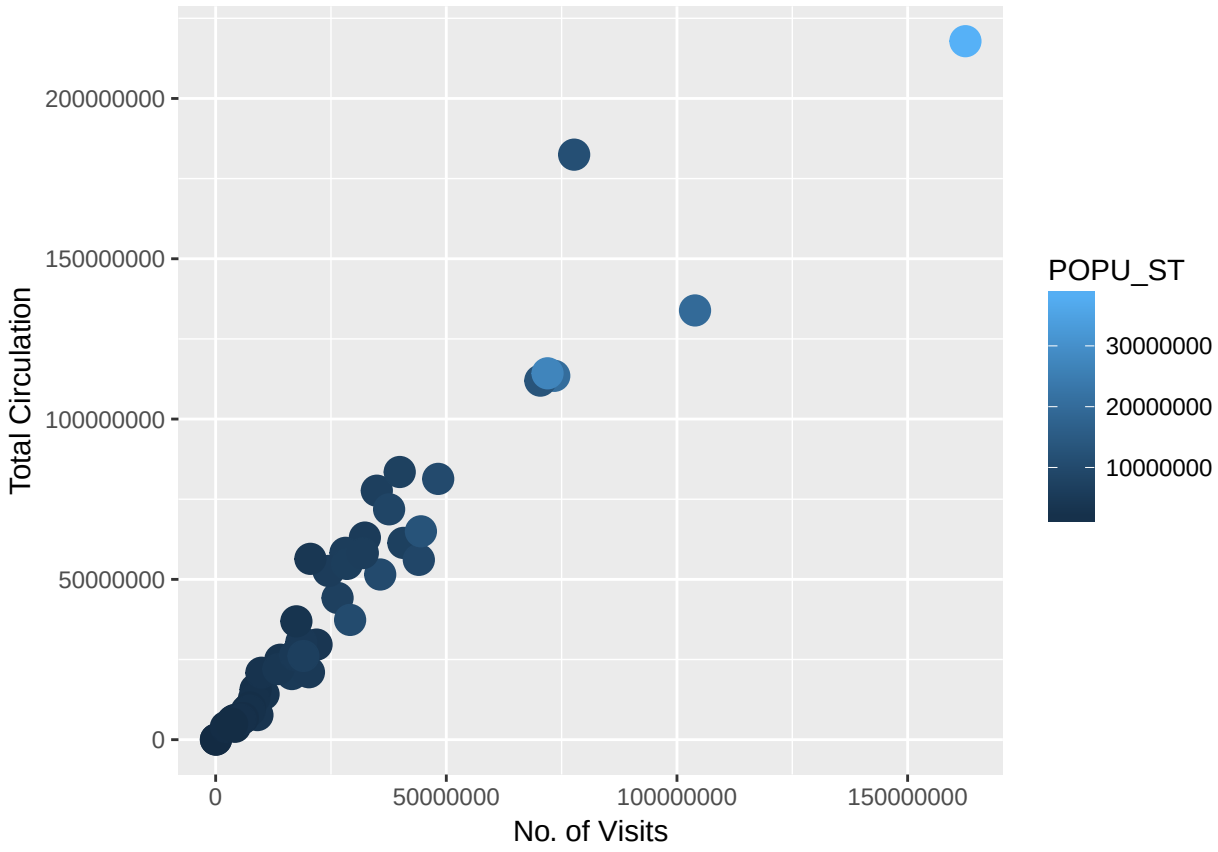
These plots examine the relative numbers of total circulation (which includes the other two variables), circulation of kid's books and circulation of electronic materials.



This scatterplot shows the effects of population size on total circulation numbers, colored by number of visits. The outlier with a relatively low population and second-highest circulation number is Ohio.



Effects of population size on number of visits, colored by total circulation numbers. As population size increases, so do number of visits.



This plot shows the relationship between number of visits and total circulation with circulation rising as visits go up. As with the plot above, the population numbers do not map perfectly cleanly.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

I was particularly interested to see that hours open and number of programs both influenced number of visits, with hours open having a greater influence. Though not exactly surprising (more people will do things at libraries that aren't programs, than people who will attend programs), it's nice to have it broken down.

Were there any interesting or surprising interactions between features?

Unsurprisingly, circulation is the action most done of the variables examined, and circulation of children's materials always outpaces electronic circulation. As Americans increase connectivity, I would be very interested to see how that relationship changes, if it ever does.

One thing I was really pleased to learn was that although it was occasionally annoying to only have 53 data points for any variable, this makes outliers easy to identify, and set up as case studies or potential deep dives into state-level policies or data that could influence their outlier-ness.

I was really surprised to see the precise, consistent relationship between loans to and loans from. It raises the question if there's some kind of policy that means the two must equal each other, or does it really naturally work out that way?

There's also a really fun outlier when data from both tables is compared, and is visible on scatter and barplots. Ohio has unusually high circulation numbers for its population and number of visits, and I would love to know what causes that!

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

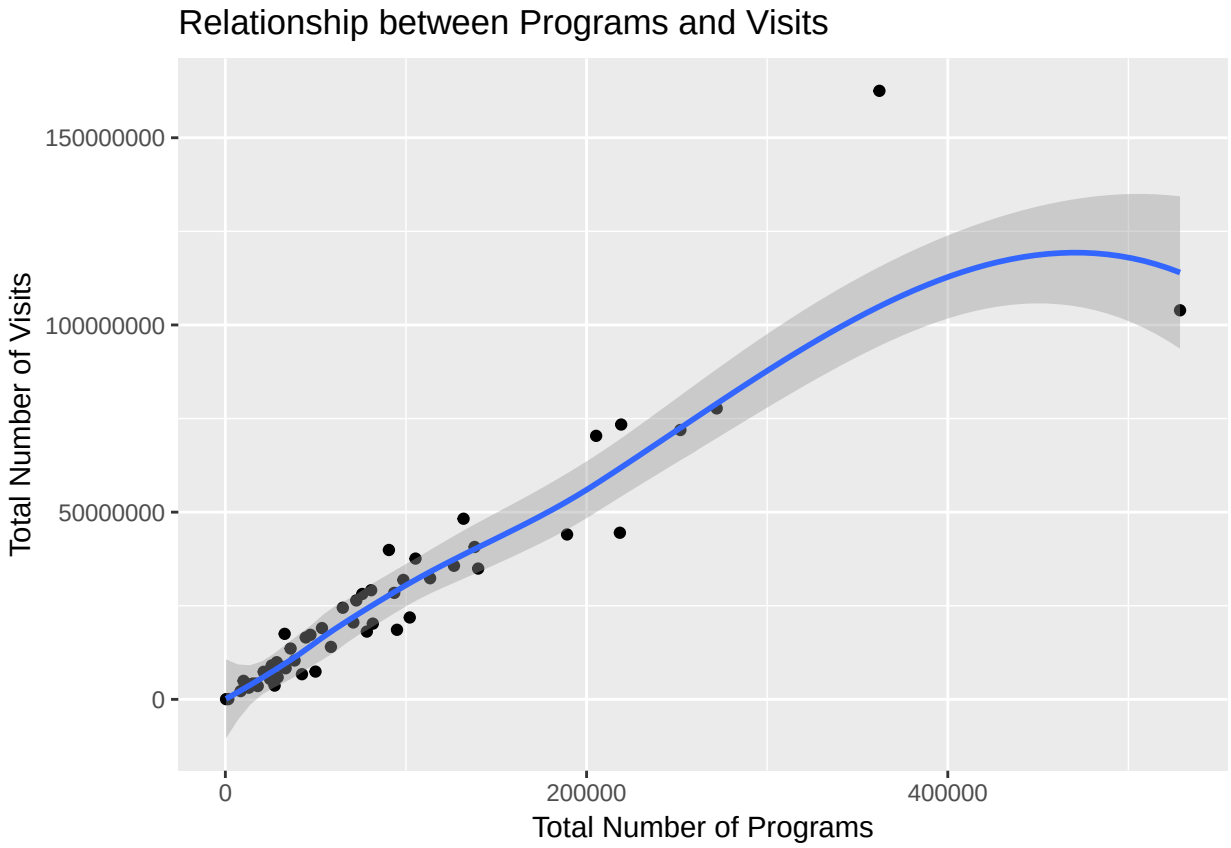
I created several linear regression models. The one that models the effects of hours open plus number of programs on visit numbers is particularly interesting to me, since it gives a better idea of what is the best predictor.

The biggest limitation is that I really am looking at a very small number of variables, especially for something like visit numbers. There are so many influences on that, that looking at one or two is more valuable for comparing them to each other, than seriously assuming that they have an effect on the feature I'm modeling.

For example, visit numbers are likely driven in part by variables not even located in these tables – access to libraries, locations of populations, awareness of the library resources, general feelings of welcome, etc.

Final Plots and Summary

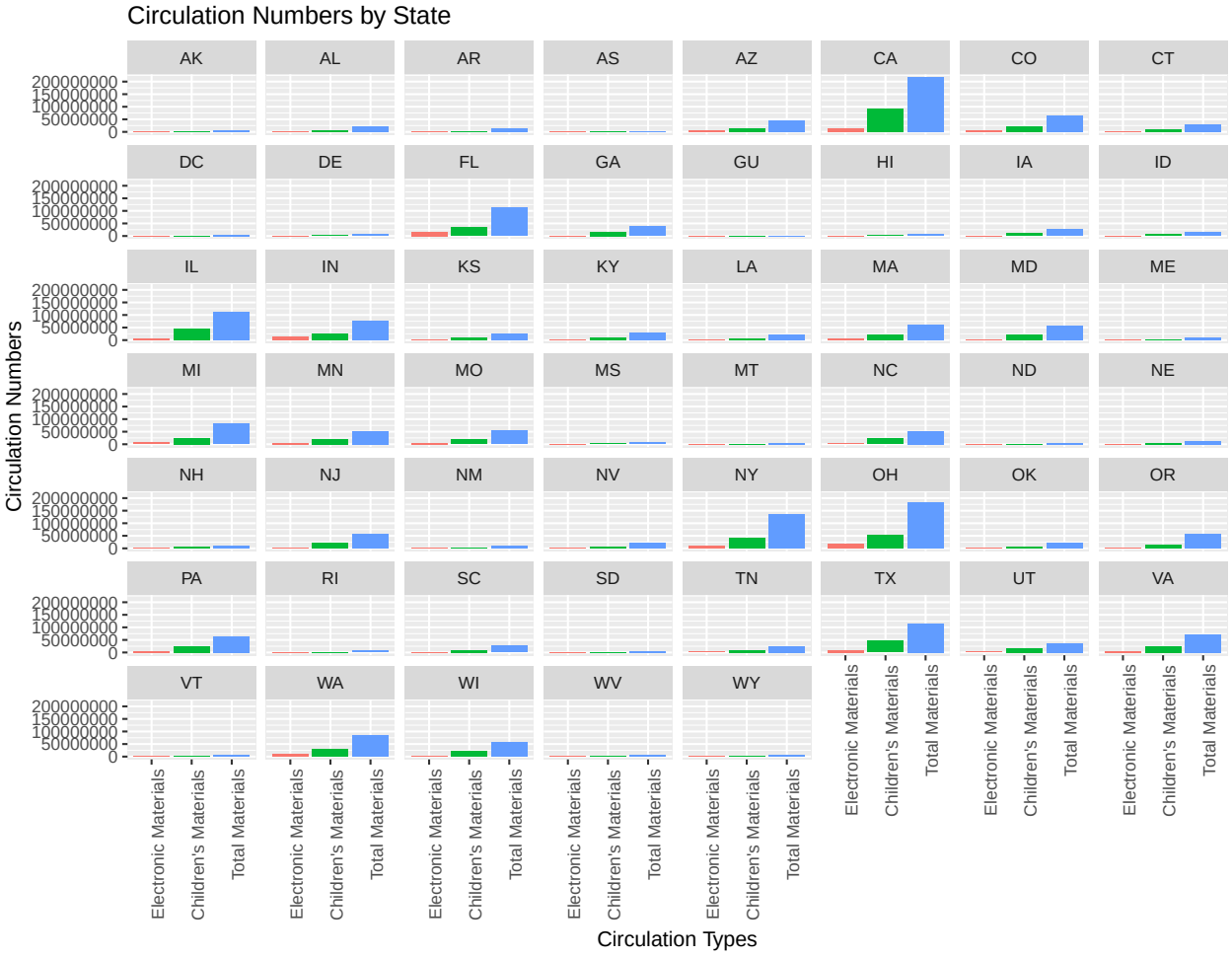
Plot One



Description One

I chose this plot because it's an excellent representative of most of the relationships I examined – largely linear, with a few interesting outliers. With the smoothing, it's clearer in this plot what was discovered in the linear modeling: that library programs as a predictor variable does not have a completely perfect relationship to visit numbers. That said, when combined with hours open, program numbers is still significant to a p-value of .01, and when assessed alone, its adjusted r-squared value is .8377, a relationship first established by and confirmed by this plot.

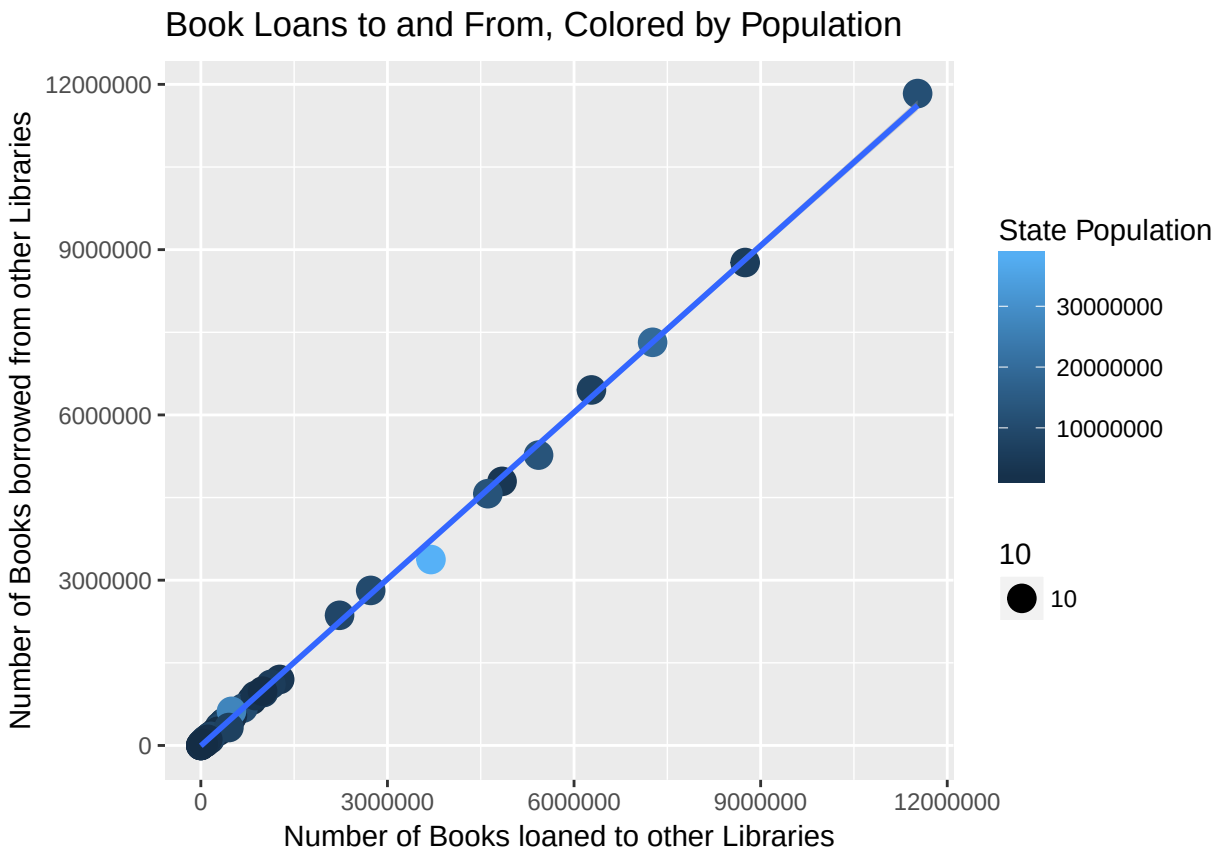
Plot Two



Description Two

I chose this plot as it gives a good overview of the way circulation differs from state to state, and lets one pull out the most interesting data at the state level. It also shows up the limitations of the dataset – the radically smaller numbers in AS and GU mean that they don't actually offer useful information at this scale. It also makes clear that circulation of children's materials always outpaces that of electronic materials, and that often children's materials make up a very significant percentage of overall circulation.

Plot Three



Description Three

I chose this plot because of how eerily regular the linear relationship between loans to and loans from is. The linear regression line goes through nearly every point, suggesting that there is probably an artificial influence on the loan system. I was also interested by the fact that population does not map onto the dataset in the same regular way – states with very high populations don't necessarily see more loans, and the state with the highest number of loans has one of the lowest populations. This (lack of) strong relationship was illustrated in the correlation plot at the beginning of the Bivariate Plots section.

A quick regression calculation, with population predicting loans to other libraries resulted in a low adjusted R-squared of .1318, confirming the lack of relationship.

Reflection

The most challenging thing was looking at the entire dataset and figuring out what questions I wanted to ask. With 124 variables to pick from, there were numerous directions I could have gone in. I chose visitor engagement because I have the most knowledge about that firsthand, having previously worked in the sector. I was able to eventually combine my two dataframes, but to some extent, I wish I had organized them better from the start.

I was really pleased at how clear the relationships were between the variables I examined – I feel spoiled by how easy the dataset was, ultimately, to work with! It was really interesting to see the linear models, and it

gave me a lot of ideas of where else this could go, even just with this variable set – for example, the effect of population size. Or, bringing in other variables, the effects of funding, number of librarians, even collection size.

I'm really interested in the few outliers in the data – why does Ohio have such a high circulation rate? Why do states with high populations have relatively low loan rates? A lot of the more unusual data was surprising. Also, having worked in the sector, I was surprised and pleased to see the relationship between number of programs and visits – when you're desperate to get people into your site, it doesn't always feel like programs do very much.