Final Project EDA

Fall 2025

```r
libraries <- read_csv("../data/libraries.csv")
```

```
## Rows: 6832 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): state, Locale
## dbl (26): Service Area Population, Total Circulation, Percentage of Children...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
numeric_libaries <- libraries %>%
  select(where(is.numeric))
full_correlation_matrix <- cor(numeric_libaries)
```

Looking at the correlation plot, we see that there are many variables, with some of them having very little correlation to our variable of interest, Total Circulation. Let's drop some.

```r
y_cor_matrix<- full_correlation_matrix[, "Total Circulation", drop = FALSE]
y_cor_matrix[order(abs(y_cor_matrix[, 1]), decreasing = TRUE), , drop = FALSE]
```

```
##                                              Total Circulation
## Total Circulation                                   1.00000000
## Physical Visits                                     0.89025206
## Local Revenue ($)                                   0.85594139
## Hours/Year                                          0.83766165
## Service Area Population                             0.81046353
## Registered Users                                    0.78177823
## Branch Library                                      0.77990997
## Internet Computers                                  0.71332256
## Children's Program Attendance                       0.70207707
## Computer Uses                                       0.69813313
## Children's Programs                                 0.66688666
## Reference Transactions                              0.62098179
## Young Adult Programs                                0.56534446
## Adult Program Attendance                            0.55176687
## Adult Programs                                      0.51878890
## State Revenue ($)                                   0.49280567
## Federal Revenue ($)                                 0.36914870
## Bookmobiles                                         0.36633538
## Other Revenue ($)                                   0.32529456
## Central Library                                    -0.30214050
## General Interest Programs                           0.28204319
```

```
## Inter-library Loans from Other Library              0.27570674
## General Interest Program Attendance                 0.25989250
## Wireless Sessions                                   0.24455640
## Inter-library Loans to Other Library                0.24005347
## Percentage of Children's Material Circulation       0.04351038
```

There are too many variables for a simple correlation matrix, and looking at this list, it seems nearly all of the numeric variables have some correlation to Total Circulation.

However, this does not show us potential collinearity, for that we can fit a simple model and check the VIF:

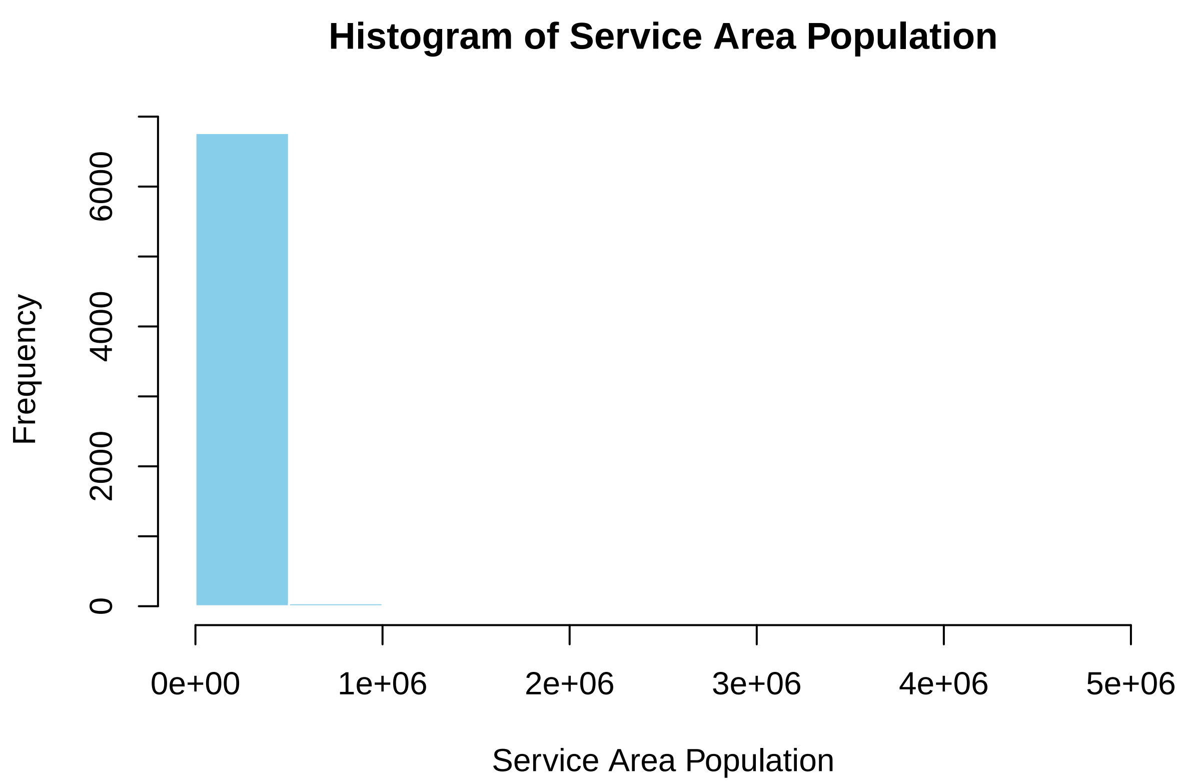
```
model <- lm(`Total Circulation` ~ ., data = numeric_libaries)
vif(model)
```

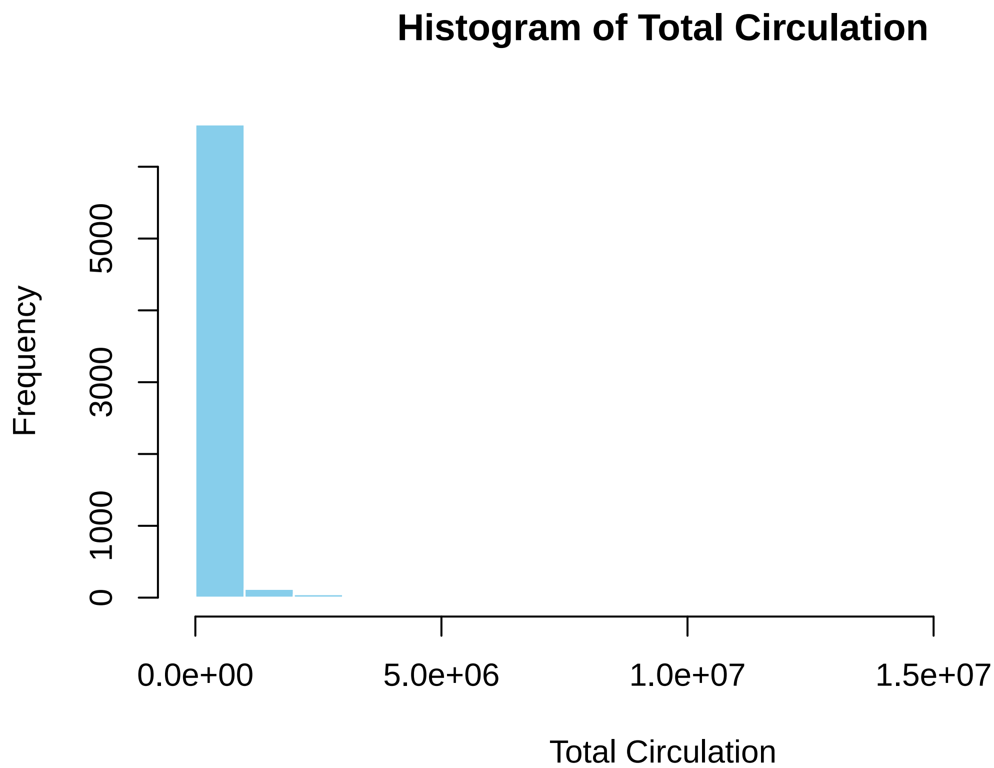
```
##                        `Service Area Population`
##                                         6.594328
## `Percentage of Children's Material Circulation`
##                                         1.034565
##                                 `Central Library`
##                                         1.183601
##                                  `Branch Library`
##                                        21.020578
##                                      Bookmobiles
##                                         1.278856
##                              `Internet Computers`
##                                         4.810188
##                                   `Computer Uses`
##                                         2.862216
##                               `Wireless Sessions`
##                                         1.088883
##                             `Children's Programs`
##                                        11.928597
##                           `Young Adult Programs`
##                                         6.529996
##                                  `Adult Programs`
##                                        11.766650
##                   `Children's Program Attendance`
##                                         8.796485
##                       `Adult Program Attendance`
##                                         6.874101
##             `General Interest Program Attendance`
##                                         1.681073
##                       `General Interest Programs`
##                                         2.413284
##                             `Local Revenue ($)`
##                                         7.826613
##                             `State Revenue ($)`
##                                         1.711105
##                           `Federal Revenue ($)`
##                                         1.745681
##                             `Other Revenue ($)`
##                                         3.957116
##                                      `Hours/Year`
##                                        31.103517
```

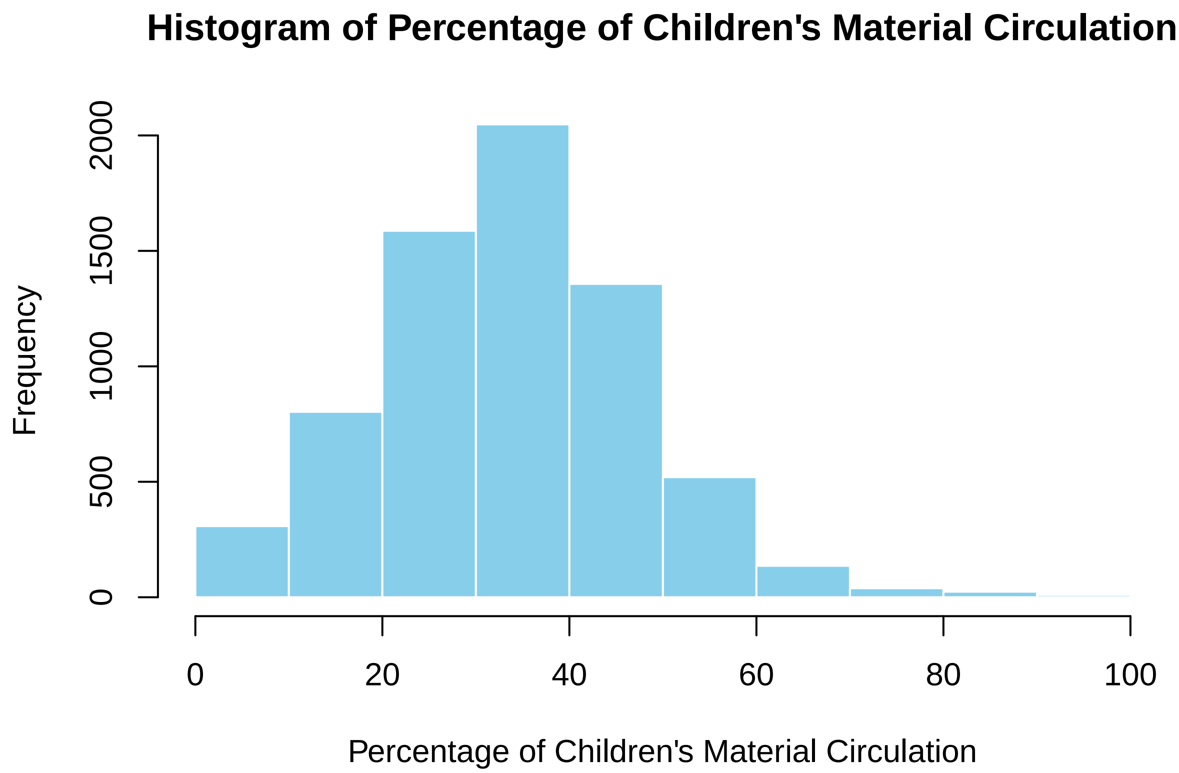
```
##                                 `Physical Visits`
##                                         15.167099
##                           `Reference Transactions`
##                                          4.545317
##                                `Registered Users`
##                                          5.614272
##              `Inter-library Loans to Other Library`
##                                          5.856800
##            `Inter-library Loans from Other Library`
##                                          6.043276
```

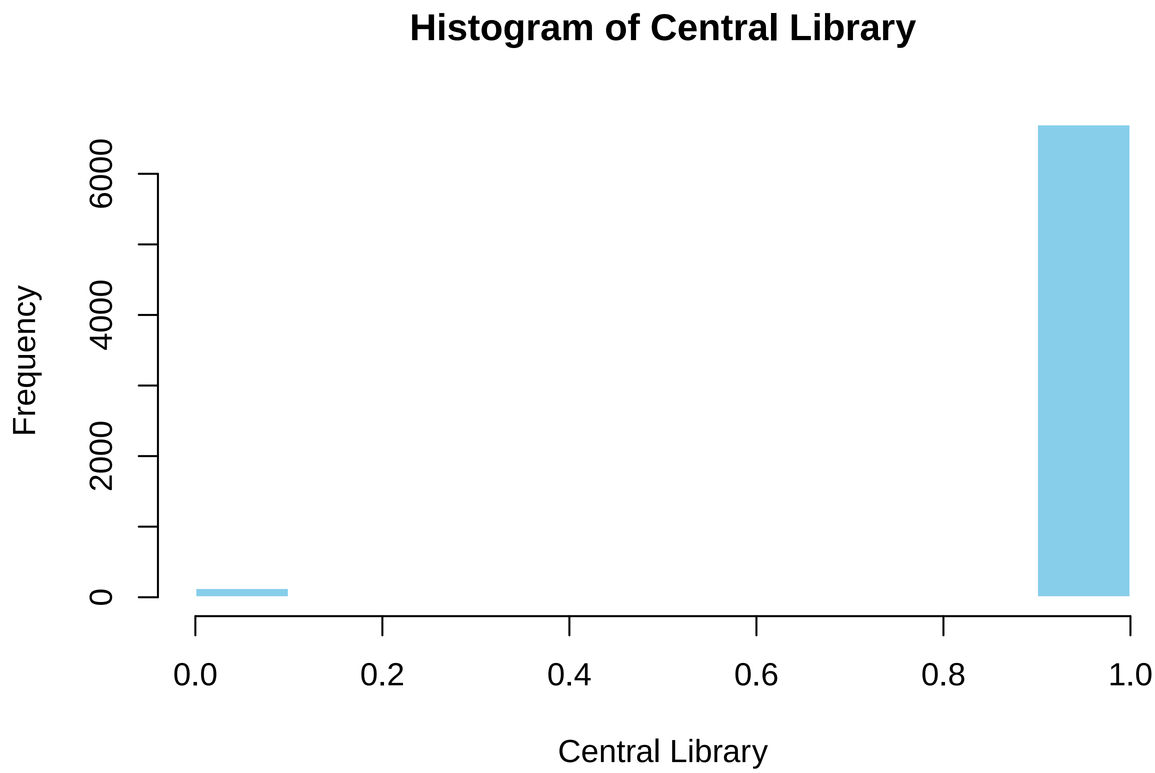
This shows that there is a generally high level of collinearity, and we should keep this in mind for the analysis.

```
for (col in names(numeric_libaries)) {
  hist(
    numeric_libaries[[col]],
    main = paste("Histogram of", col),
    xlab = col,
    col = "skyblue",
    border = "white"
  )
}
```

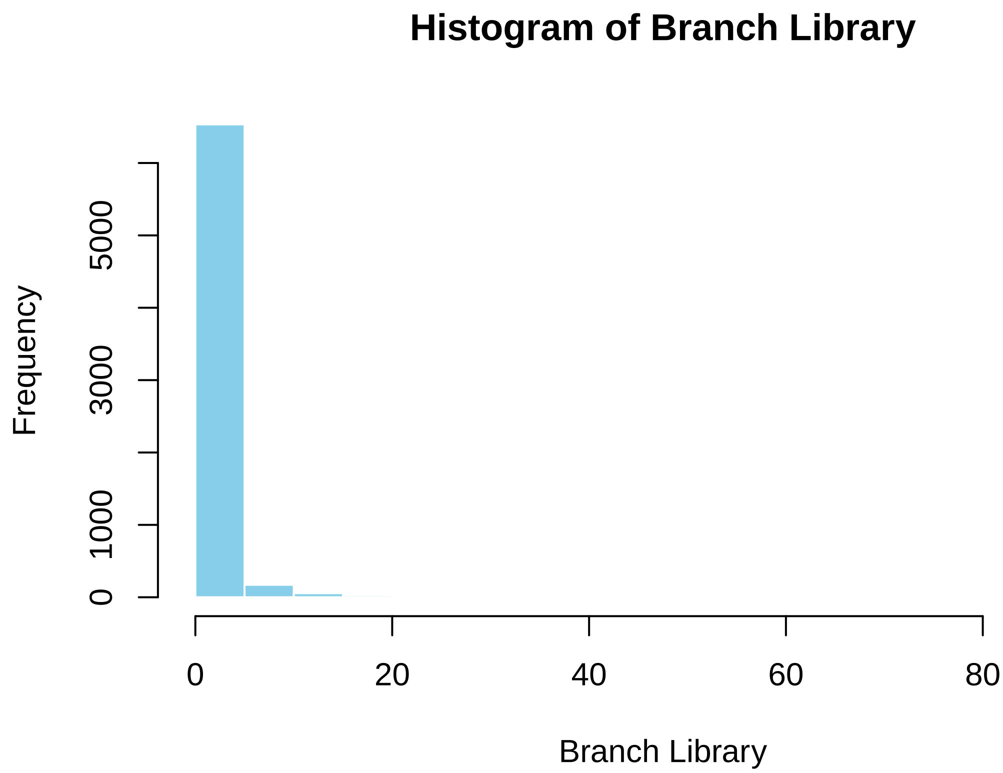



Histogram of Total Circulation
Frequency
0
1000
3000
5000
0.0e+00
5.0e+06
1.0e+07
1.5e+07
Total Circulation

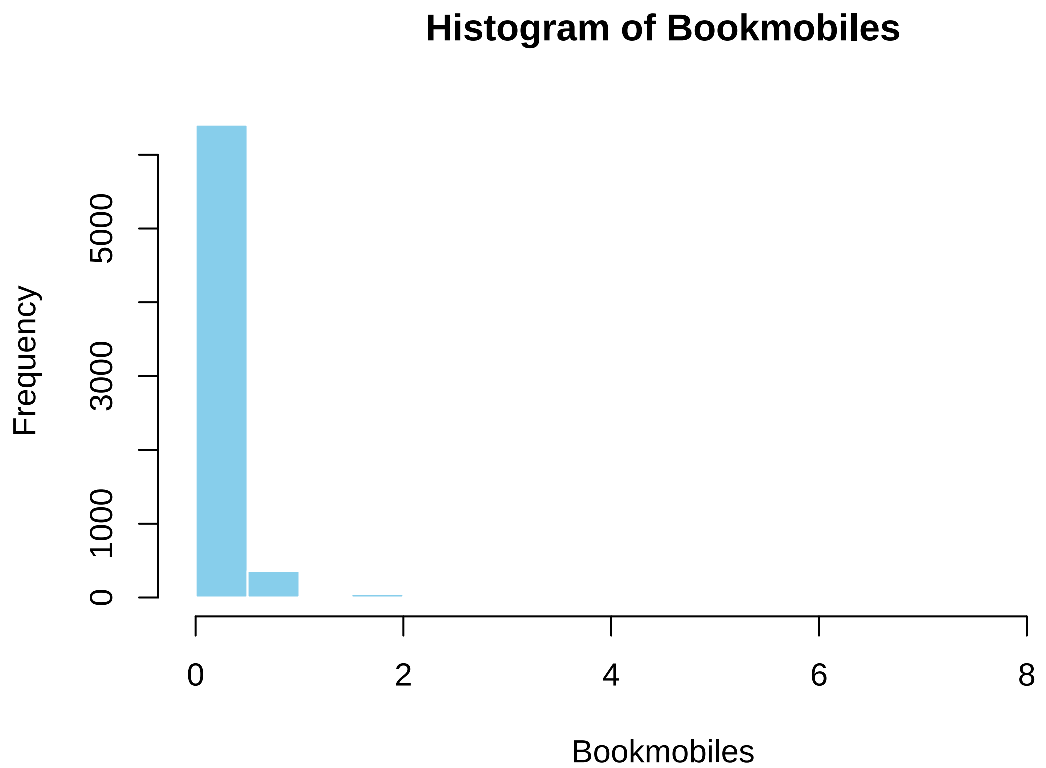

Histogram of Percentage of Children's Material Circulation
Frequency
0
500
1000
1500
2000
0
20
40
60
80
100
Percentage of Children's Material Circulation

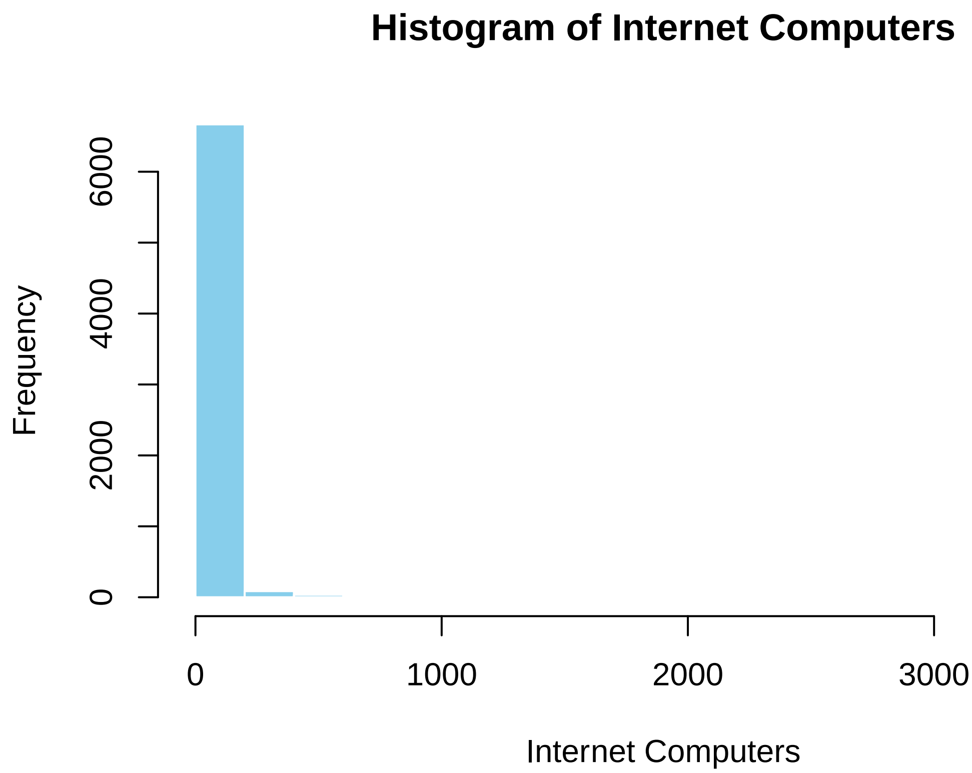


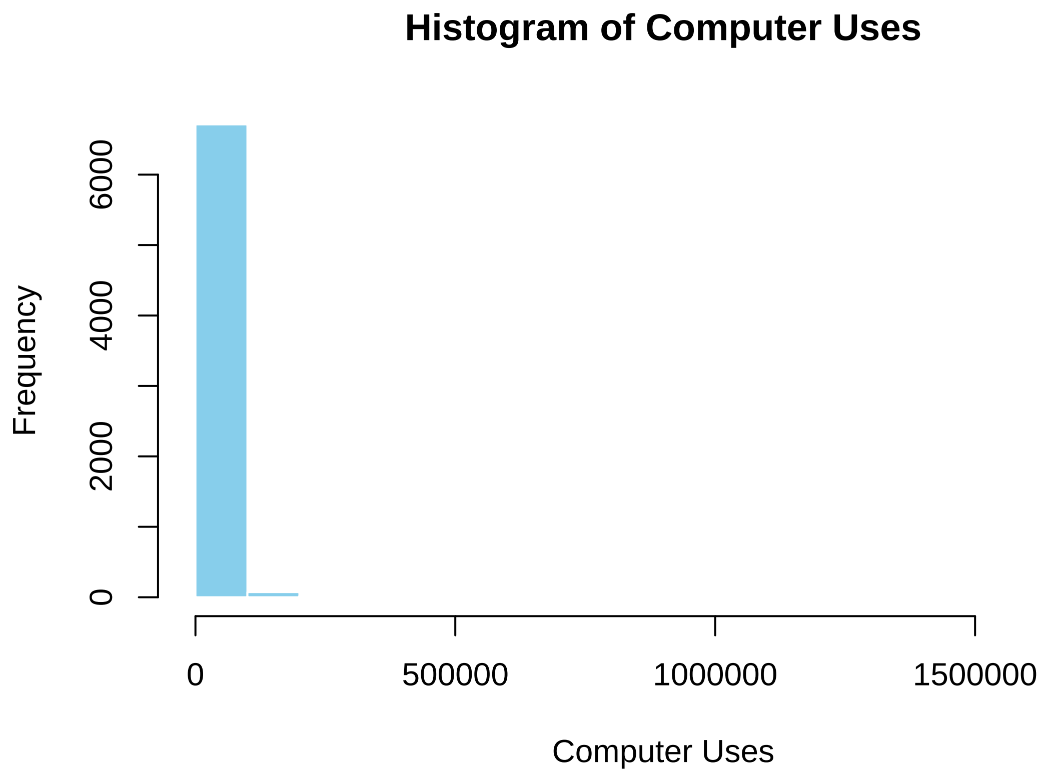
Histogram of Central Library
Frequency
0
2000
4000
6000
0.0
0.2
0.4
0.6
0.8
1.0
Central Library

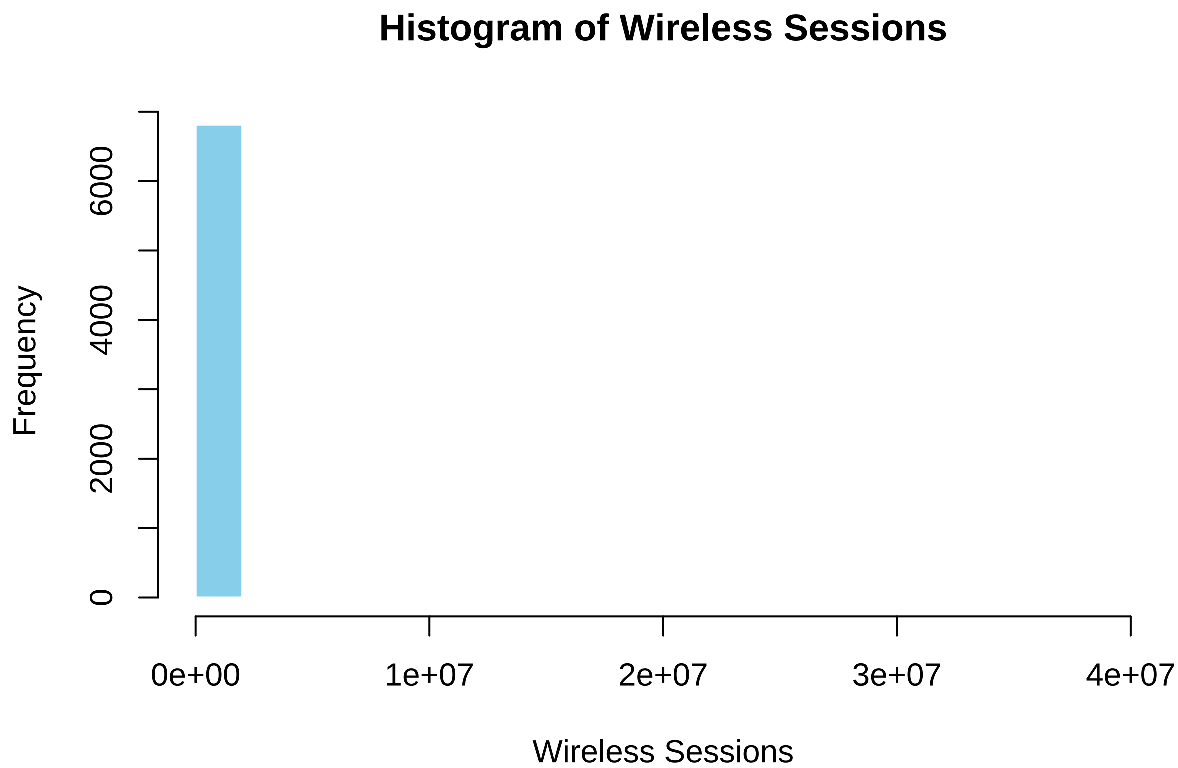


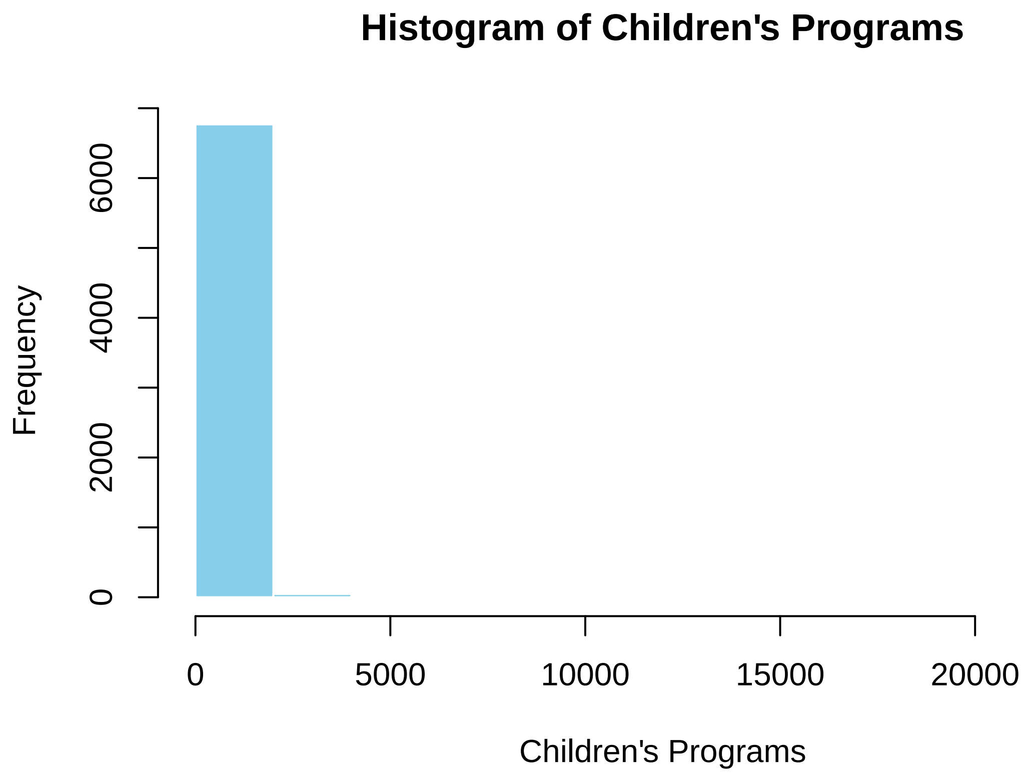
Histogram of Branch Library
Frequency
0
1000
3000
5000
0
20
40
60
80
Branch Library

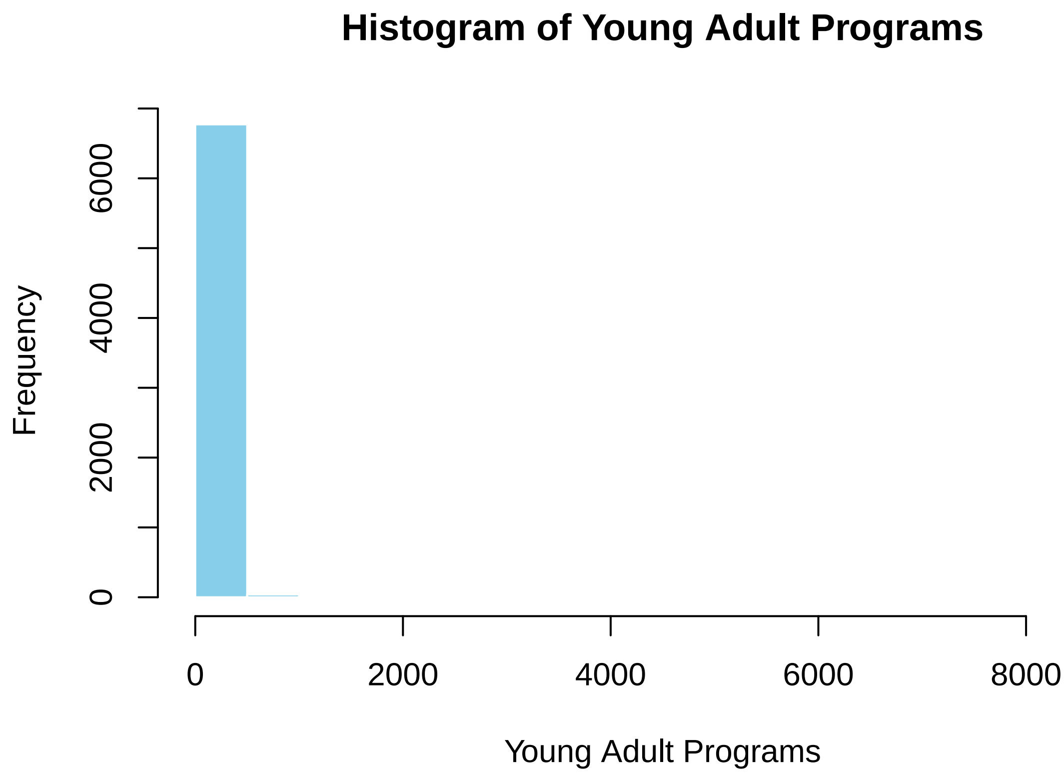

Histogram of Bookmobiles
Frequency
0
1000
3000
5000
0
2
4
6
8
Bookmobiles

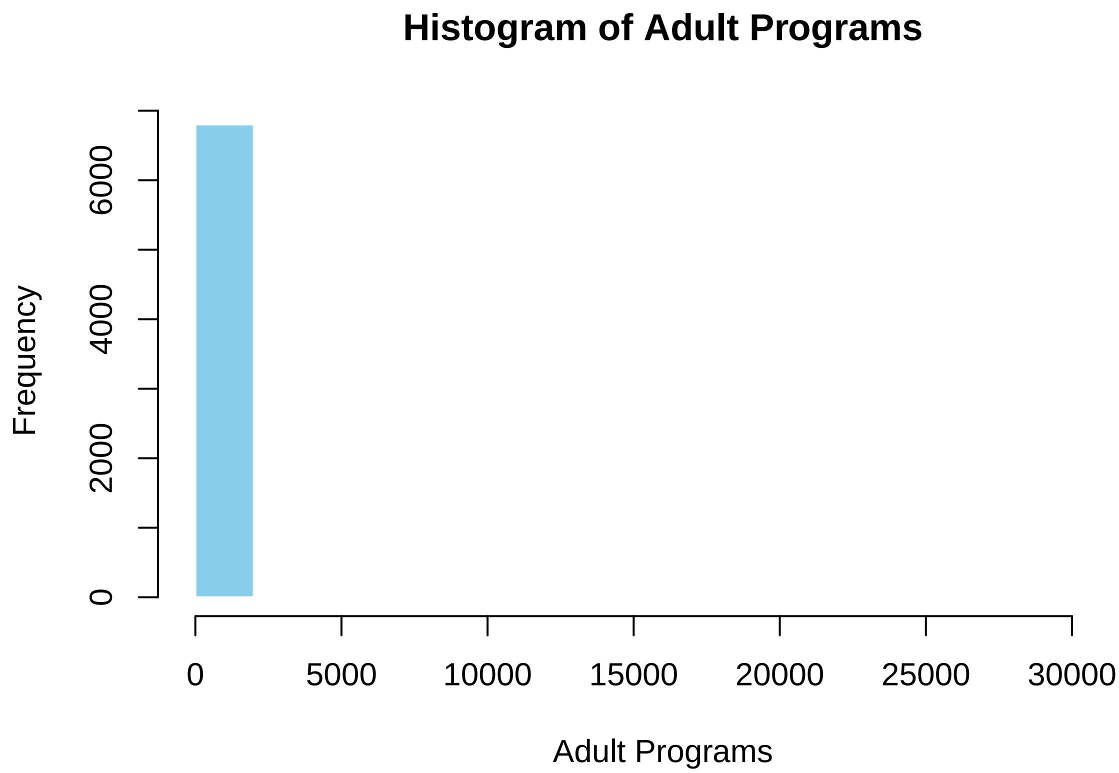

Histogram of Internet Computers
Frequency
0
2000
4000
6000
0
1000
2000
3000
Internet Computers

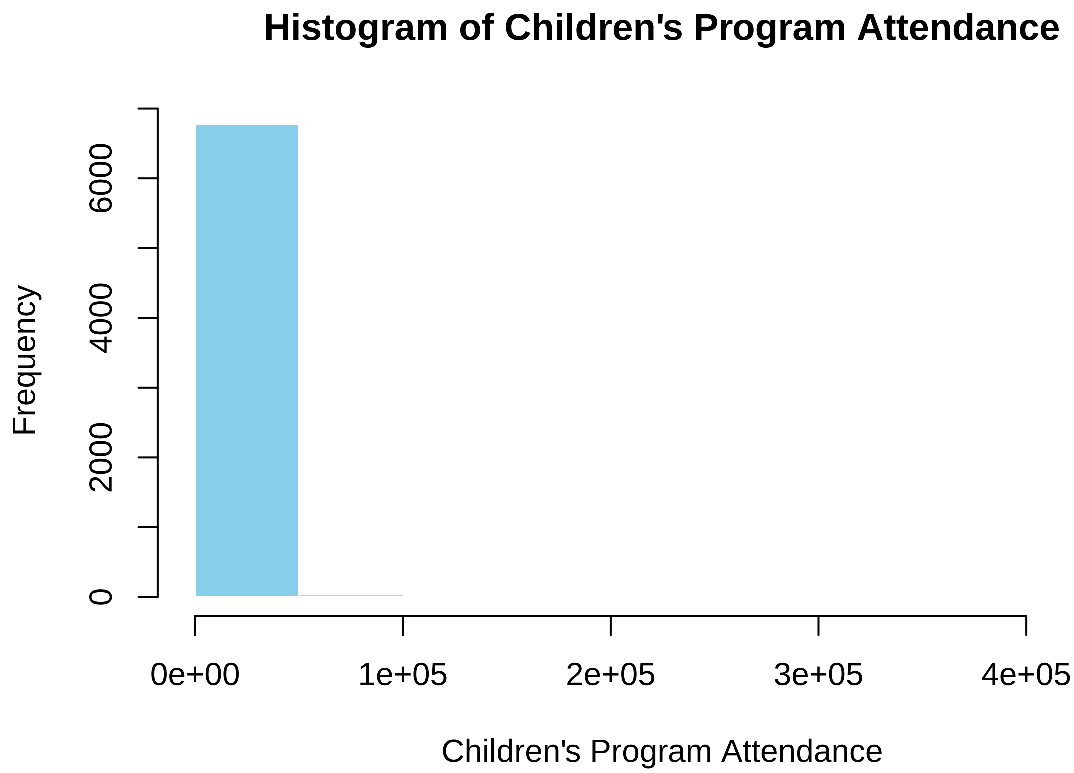

Histogram of Computer Uses
Frequency
0
2000
4000
6000
0
500000
1000000
1500000
Computer Uses

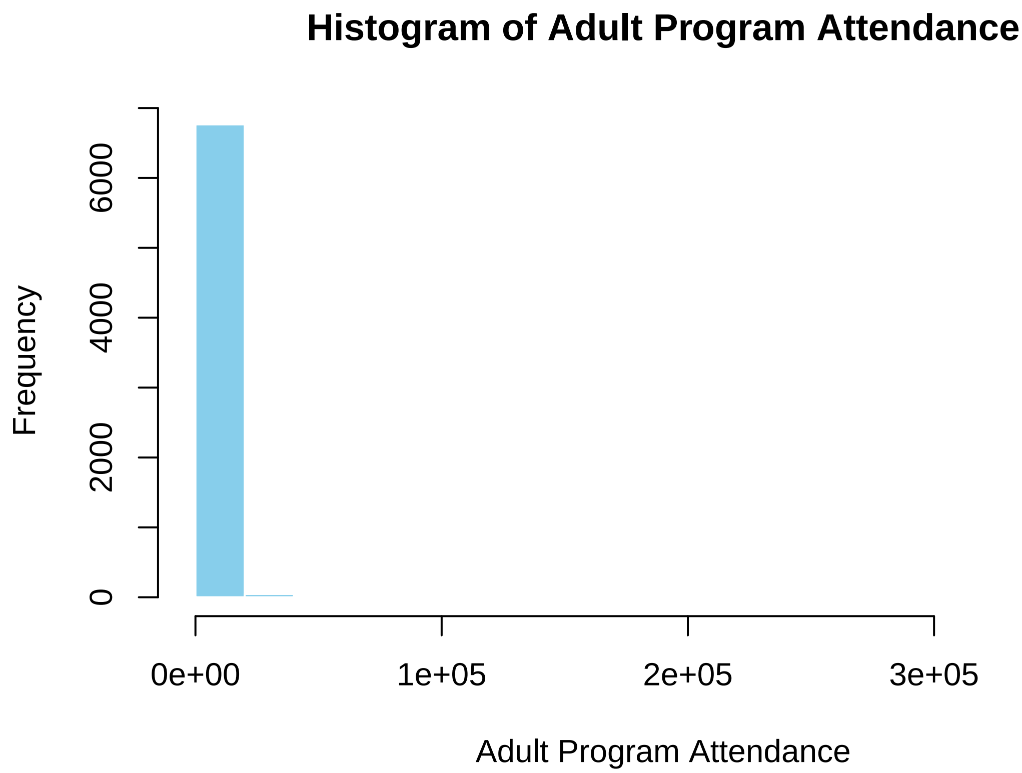

Histogram of Wireless Sessions
Frequency
0
2000
4000
6000
0e+00
1e+07
2e+07
3e+07
4e+07
Wireless Sessions

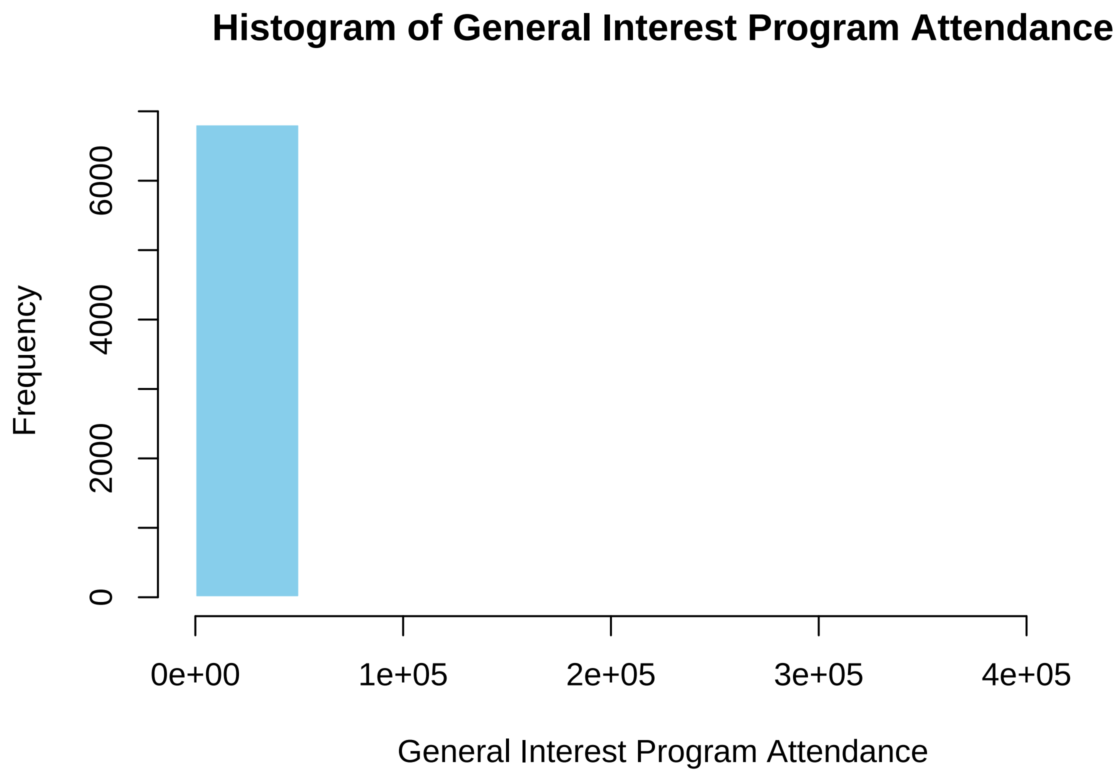

Histogram of Children's Programs
Frequency
0
2000
4000
6000
0
5000
10000
15000
20000
Children's Programs

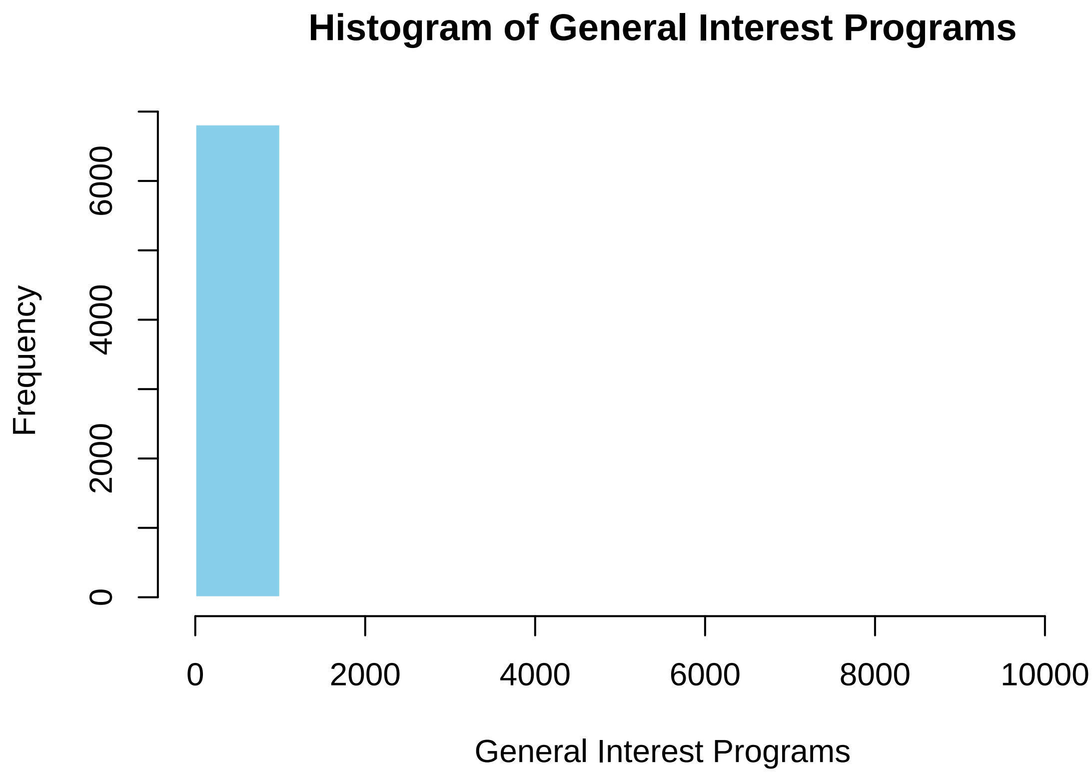

Histogram of Young Adult Programs
Frequency
0
2000
4000
6000
0
2000
4000
6000
8000
Young Adult Programs


Histogram of Adult Programs
Frequency
0
2000
4000
6000
0
5000
10000
15000
20000
25000
30000
Adult Programs

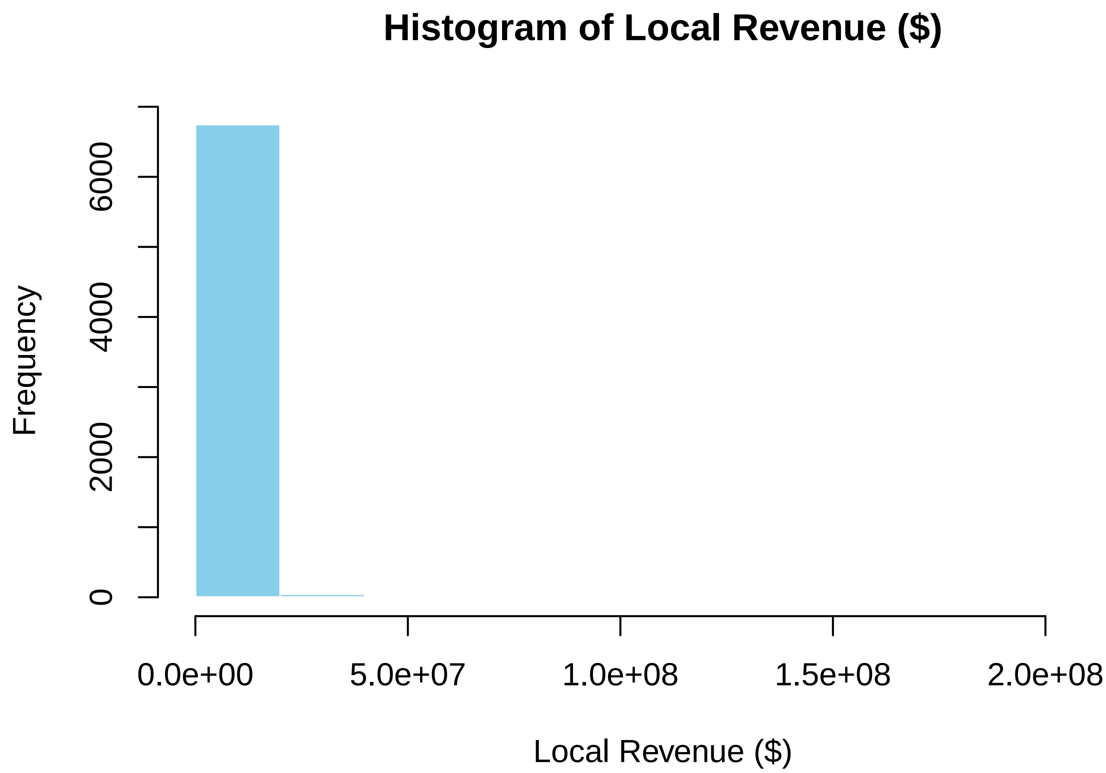

Histogram of Children's Program Attendance
Frequency
0
2000
4000
6000
0e+00
1e+05
2e+05
3e+05
4e+05
Children's Program Attendance

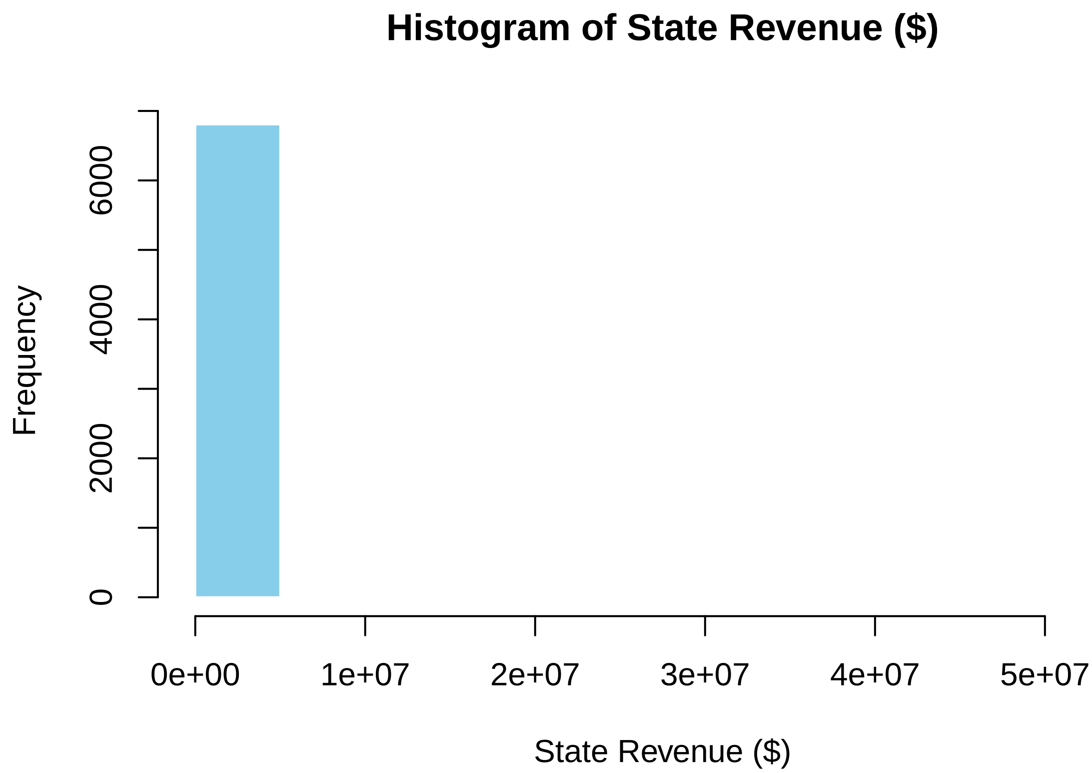

Histogram of Adult Program Attendance
Frequency
0
2000
4000
6000
0e+00
1e+05
2e+05
3e+05
Adult Program Attendance

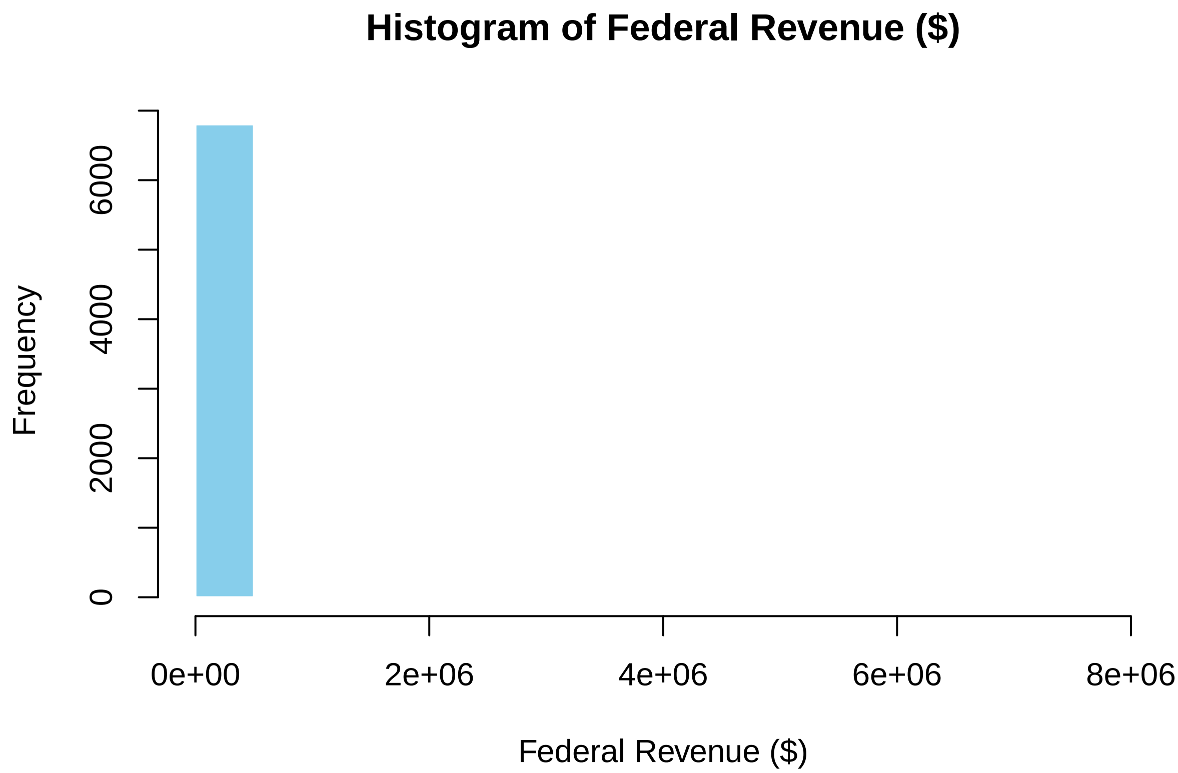

Histogram of General Interest Program Attendance
Frequency
0
2000
4000
6000
0e+00
1e+05
2e+05
3e+05
4e+05
General Interest Program Attendance

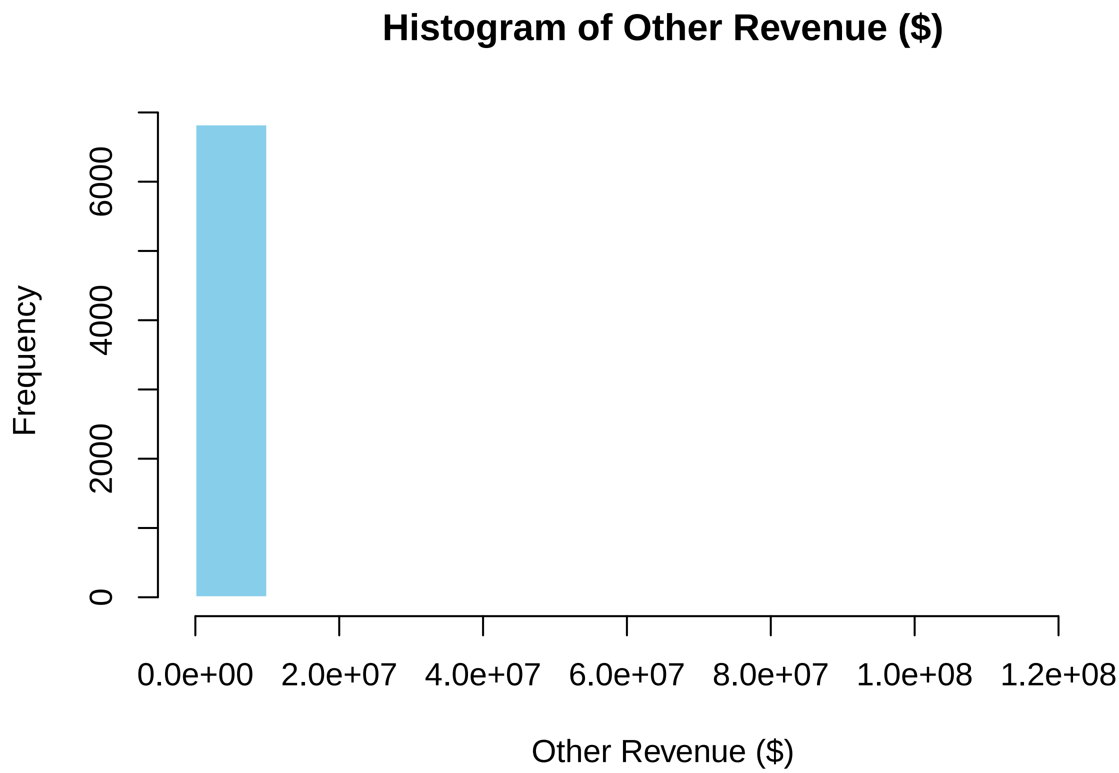


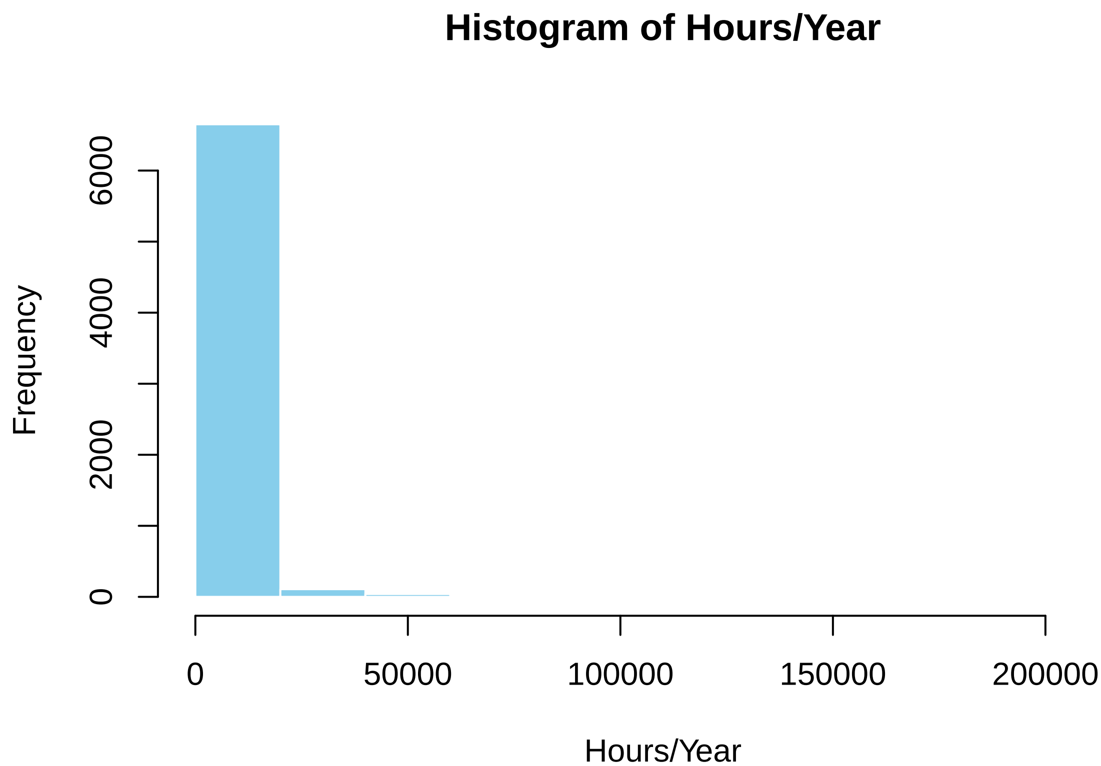
Histogram of General Interest Programs
Frequency
0
2000
4000
6000
0
2000
4000
6000
8000
10000
General Interest Programs

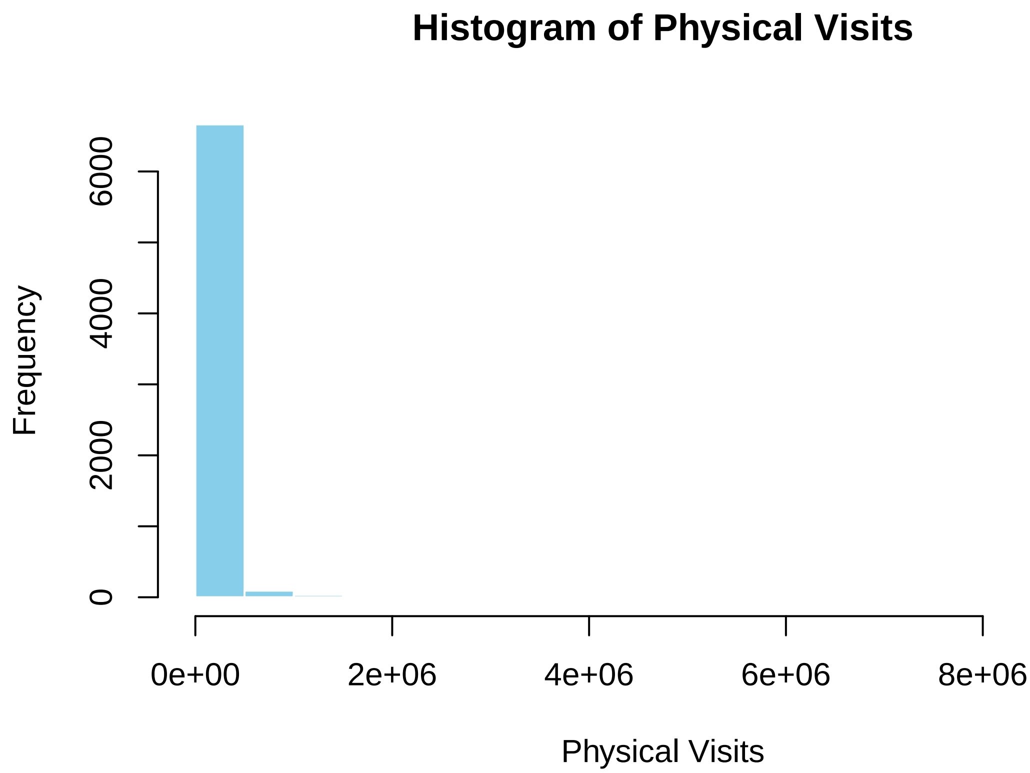

Histogram of Local Revenue ($)
Frequency
0
2000
4000
6000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
Local Revenue ($)

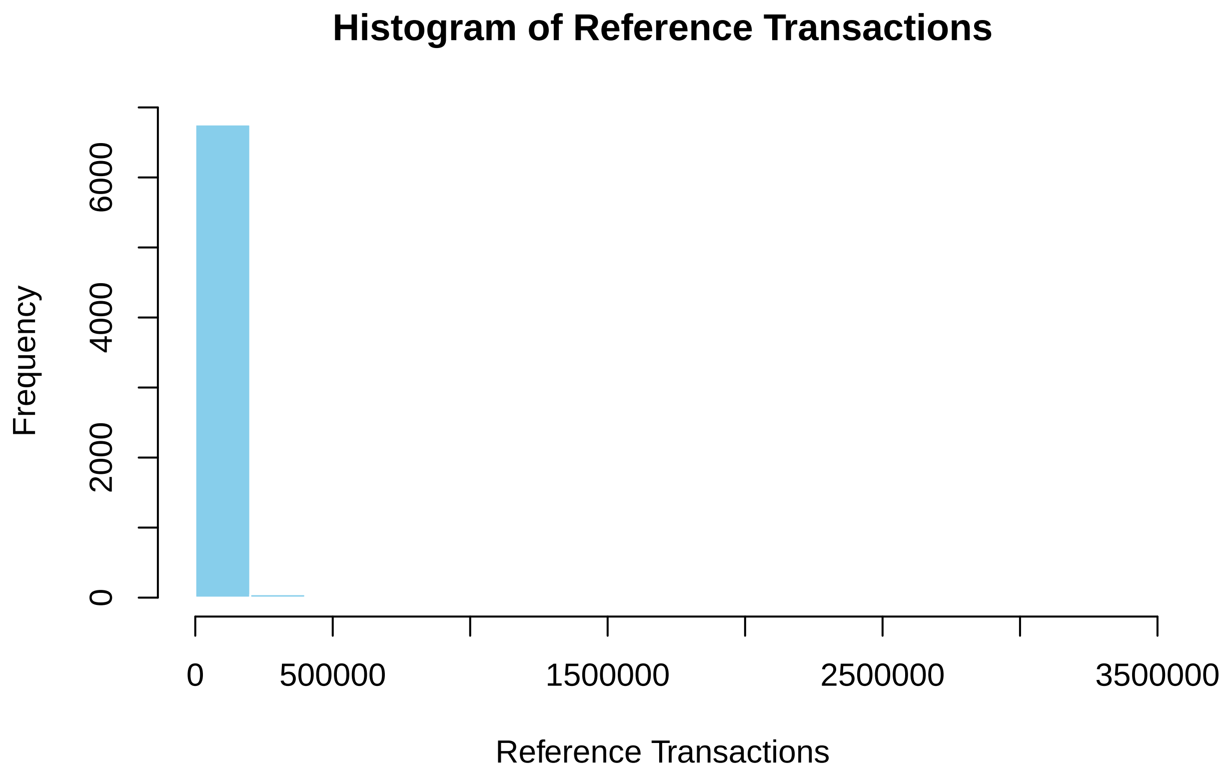

Histogram of State Revenue ($)
Frequency
0
2000
4000
6000
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
State Revenue ($)

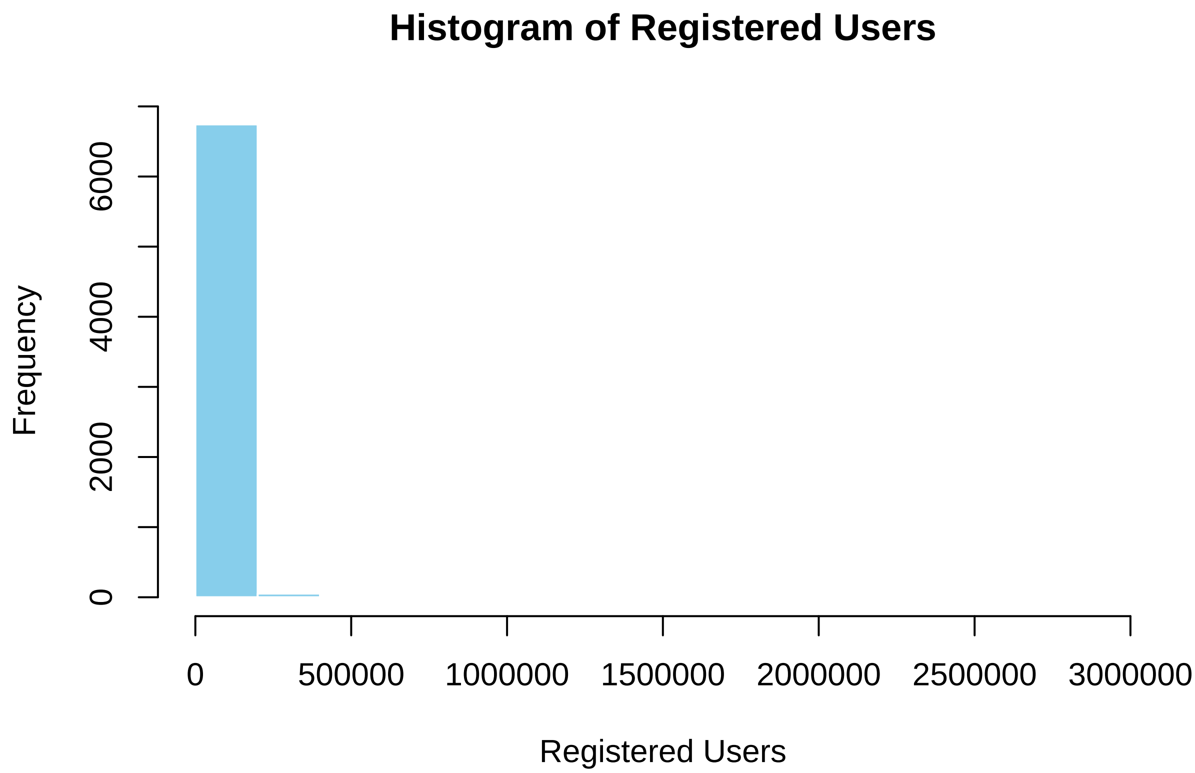

Histogram of Federal Revenue ($)
Frequency
0
2000
4000
6000
0e+00
2e+06
4e+06
6e+06
8e+06
Federal Revenue ($)

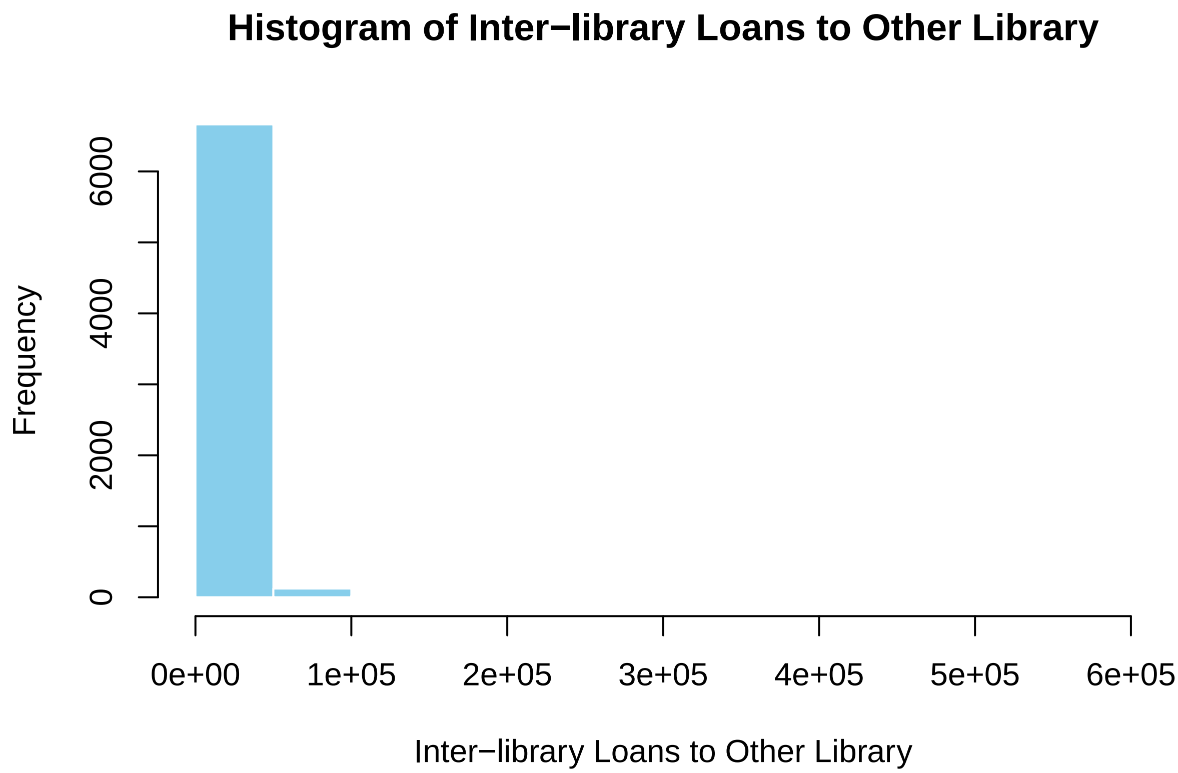

Histogram of Other Revenue ($)
Frequency
0
2000
4000
6000
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
Other Revenue ($)

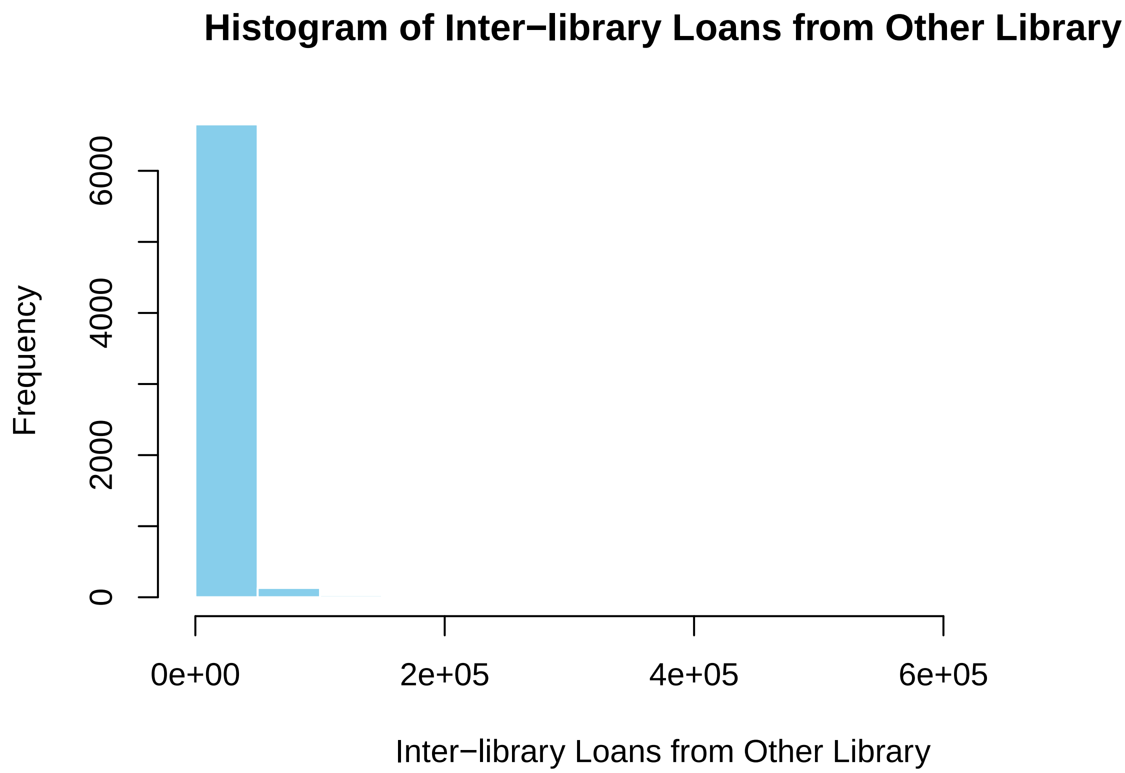

Histogram of Hours/Year
Frequency
0
2000
4000
6000
0
50000
100000
150000
200000
Hours/Year


Histogram of Physical Visits
Frequency
0
2000
4000
6000
0e+00
2e+06
4e+06
6e+06
8e+06
Physical Visits


Histogram of Reference Transactions
Frequency
0
2000
4000
6000
0
500000
1500000
2500000
3500000
Reference Transactions


Histogram of Registered Users
Frequency
0
2000
4000
6000
0
500000
1000000
1500000
2000000
2500000
3000000
Registered Users


Histogram of Inter−library Loans to Other Library
Frequency
0
2000
4000
6000
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
Inter−library Loans to Other Library

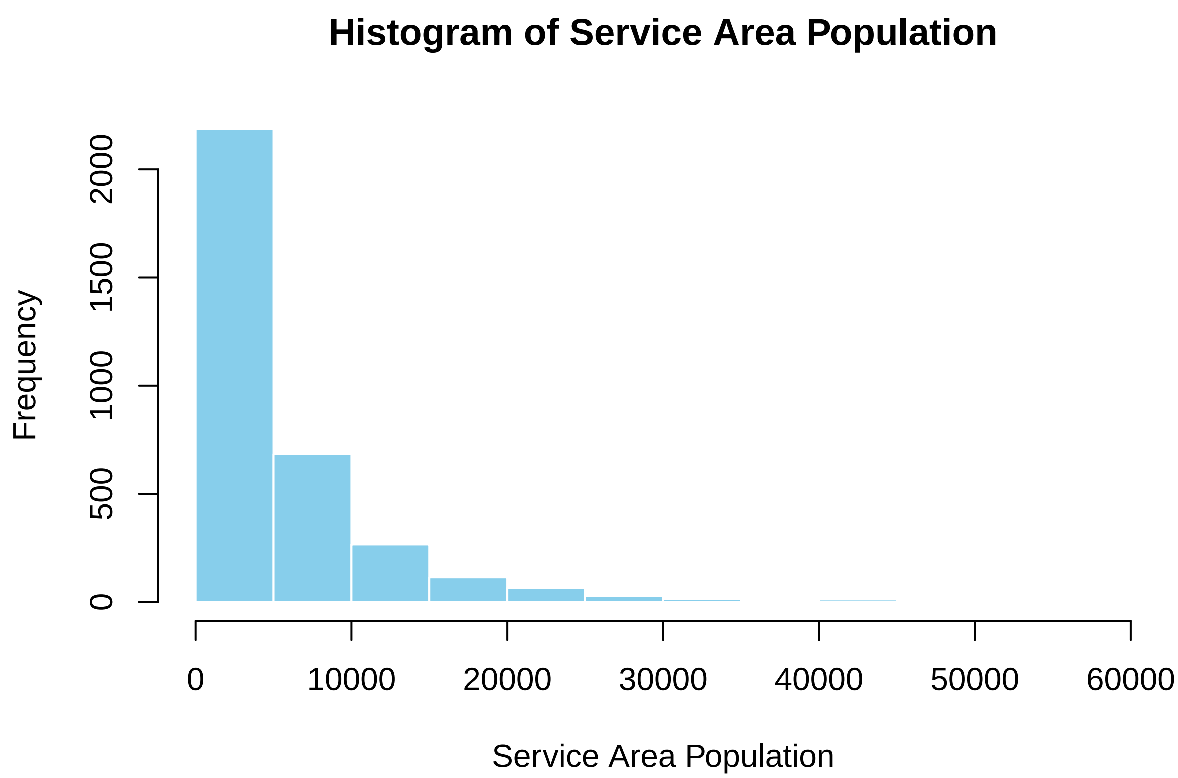


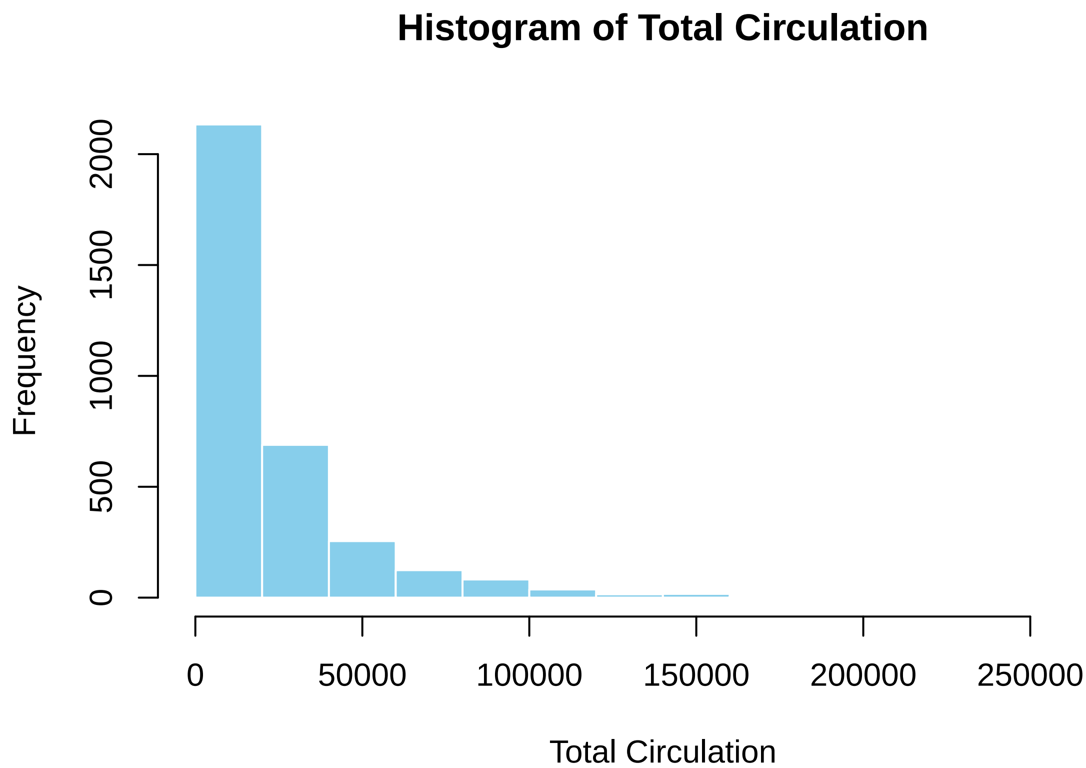
The histograms of the numeric data show that many of these have extreme outliers, so let's remove them:

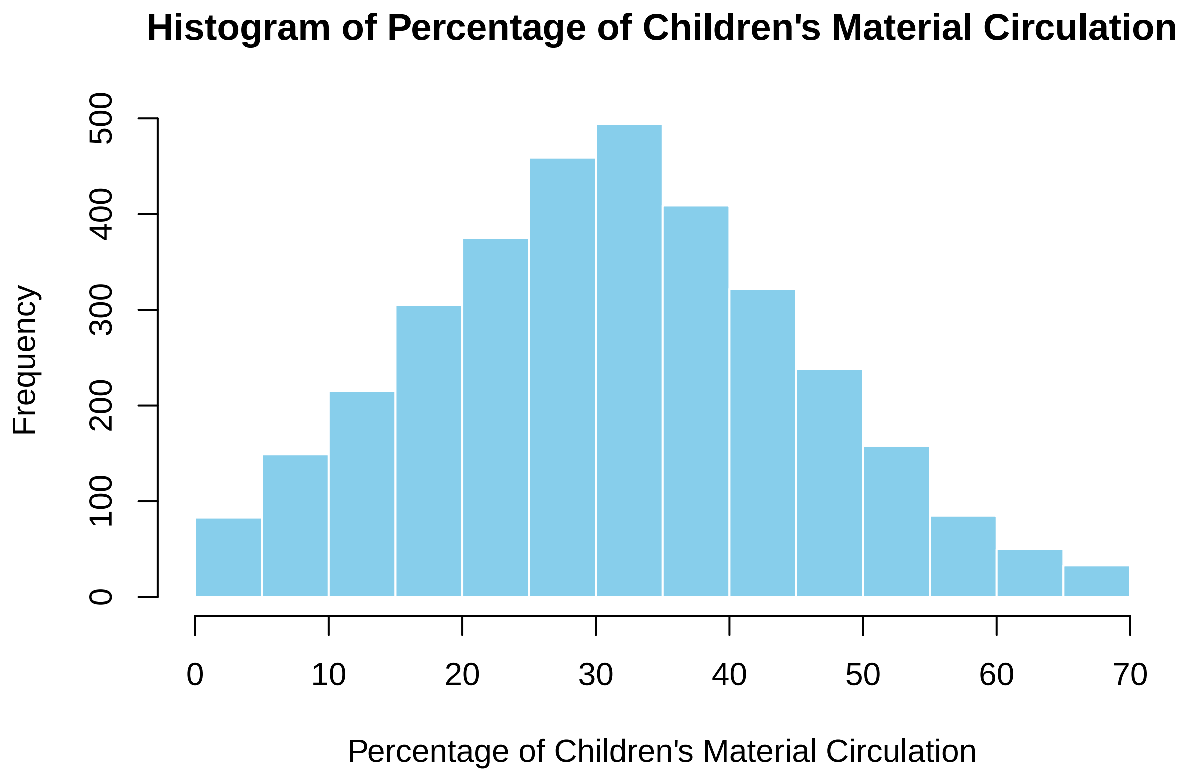
library(dplyr)

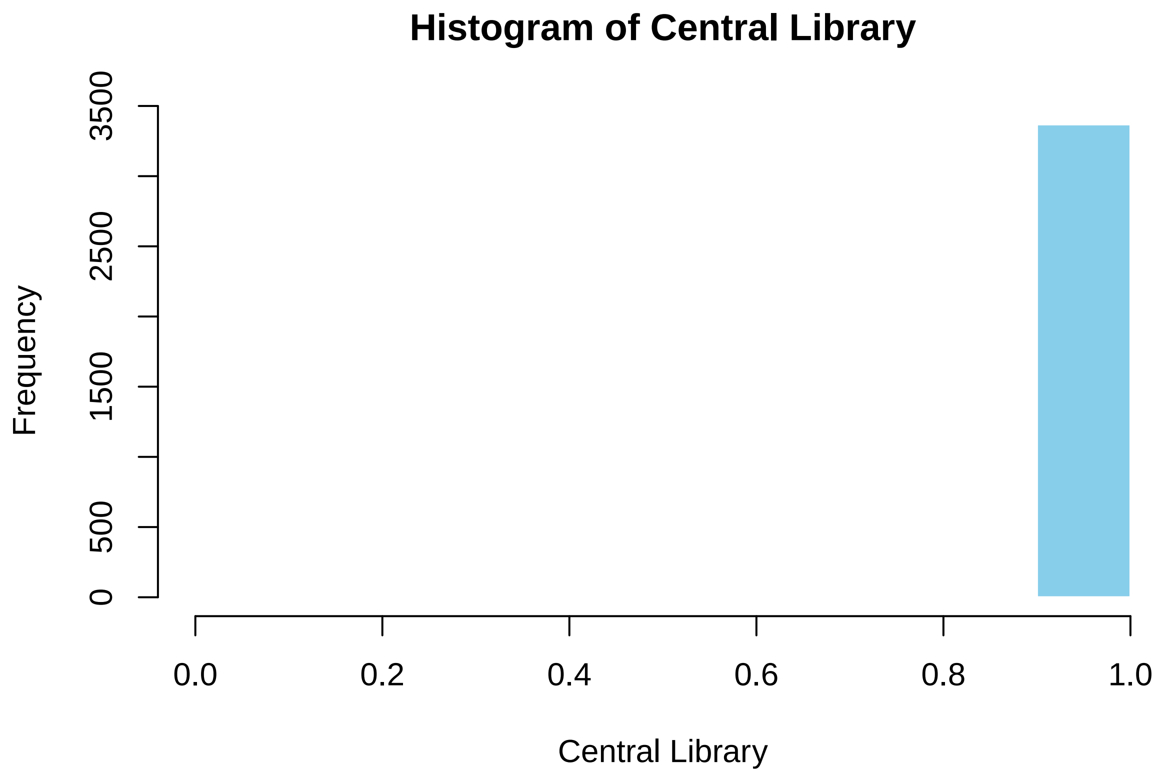
```
# Function to remove outliers using the IQR rule
remove_outliers <- function(x) {
  if (is.numeric(x) && length(unique(na.omit(x))) > 2) {  # skip binary
    q1 <- quantile(x, 0.25, na.rm = TRUE)
    q3 <- quantile(x, 0.75, na.rm = TRUE)
    iqr <- q3 - q1
    lower <- q1 - 1.5 * iqr
    upper <- q3 + 1.5 * iqr
    x[x < lower | x > upper] <- NA  # mark outliers
  }
  return(x)
}

# Columns to ignore
ignore_cols <- c("Bookmobiles", "Branch Library")

# Apply only to numeric, non-binary, and not ignored columns
libraries_no_outliers <- libraries %>%
  mutate(across(
    .cols = where(is.numeric) & !all_of(ignore_cols),
    .fns = remove_outliers
  )) %>%
  drop_na()
```

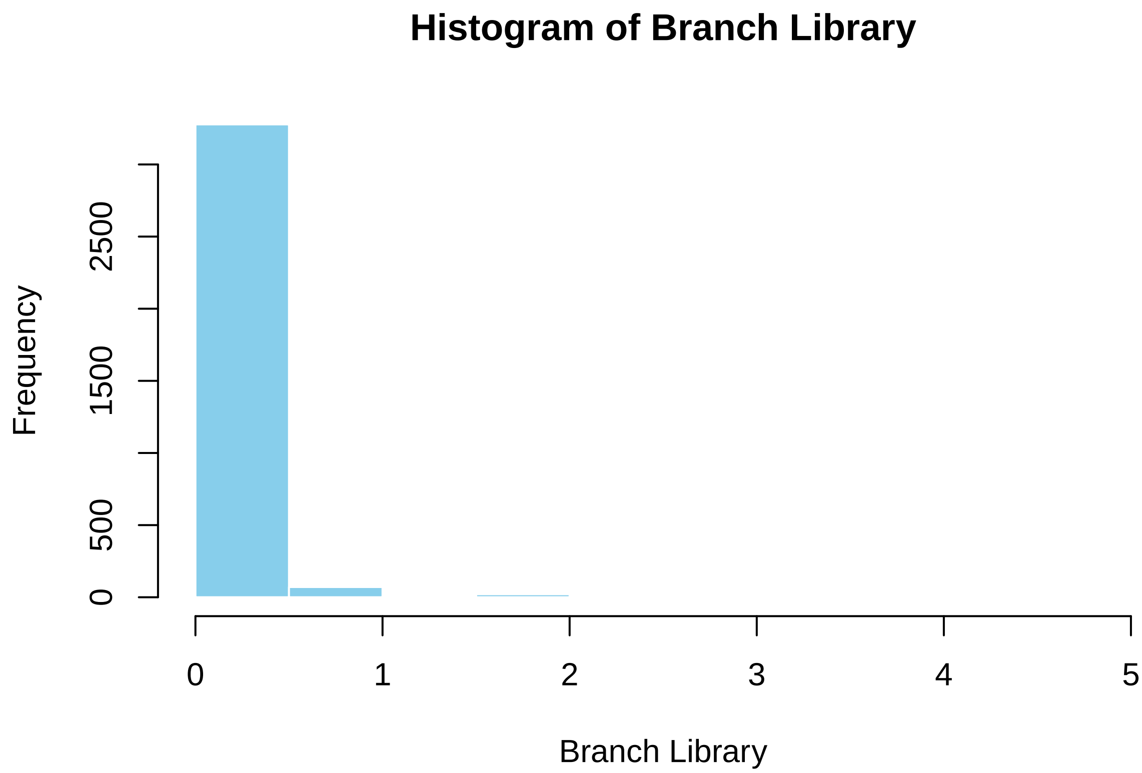
```
libraries_no_outliers_numeric <- libraries_no_outliers %>% select(where(is.numeric))
for (col in names(libraries_no_outliers_numeric)) {
  hist(
    libraries_no_outliers_numeric[[col]],
    main = paste("Histogram of", col),
    xlab = col,
    col = "skyblue",
    border = "white"
  )
}
```

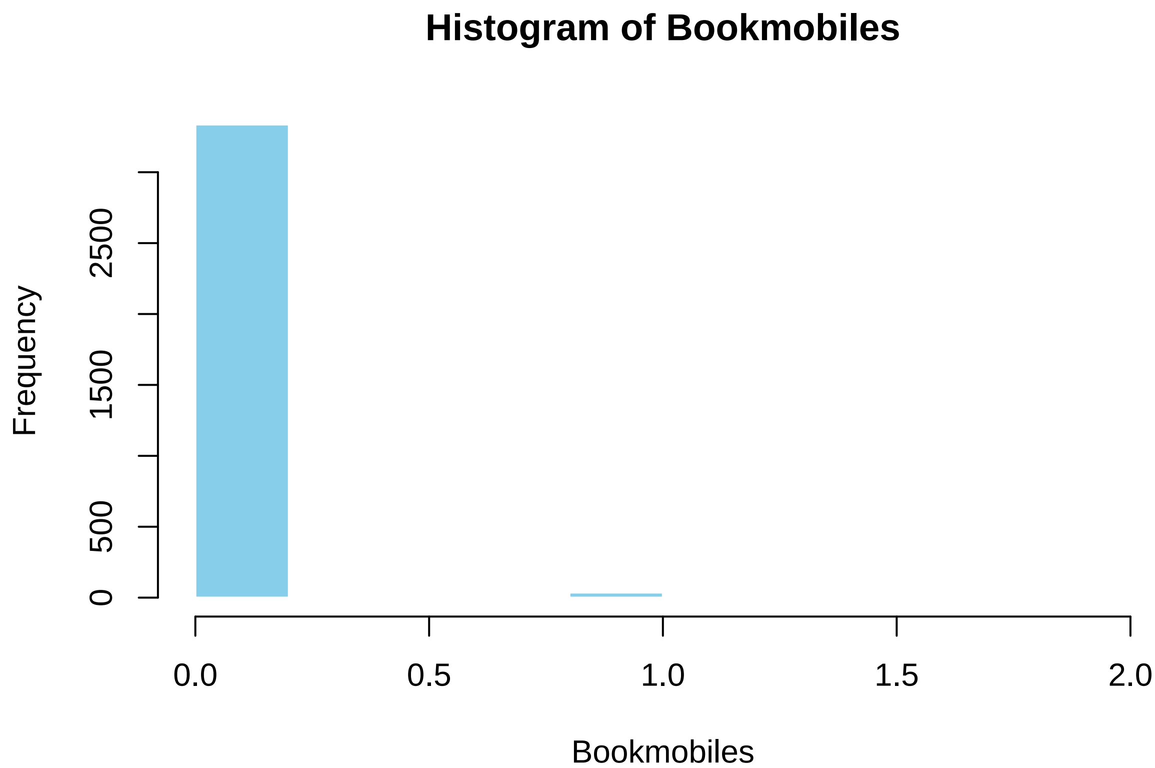


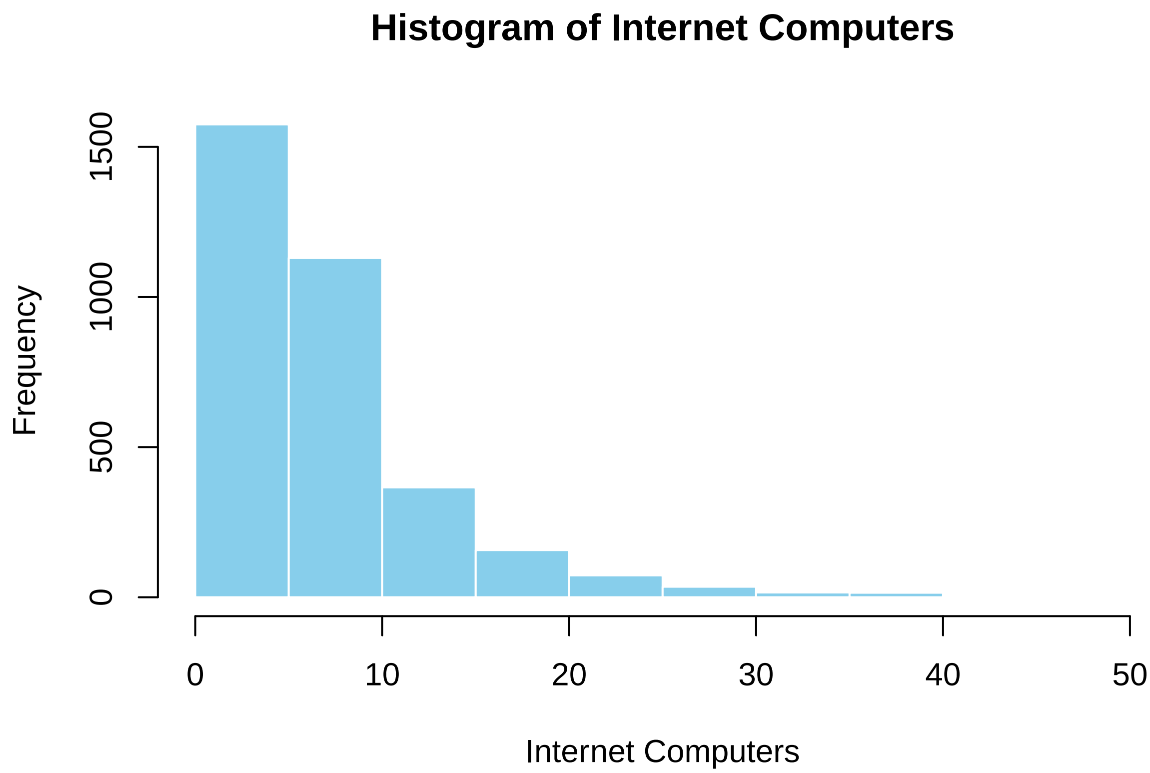

Histogram of Total Circulation
Frequency
0
500
1000
1500
2000
0
50000
100000
150000
200000
250000
Total Circulation

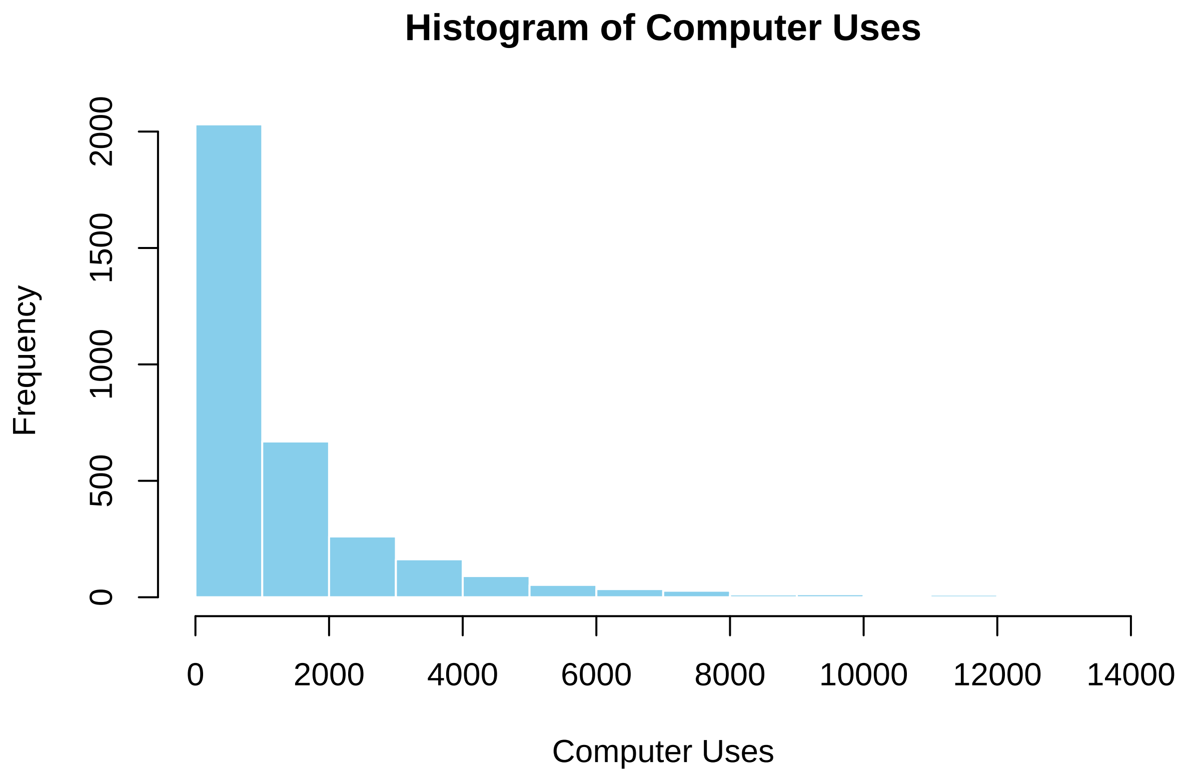

Histogram of Percentage of Children's Material Circulation
Frequency
0
100
200
300
400
500
0
10
20
30
40
50
60
70
Percentage of Children's Material Circulation

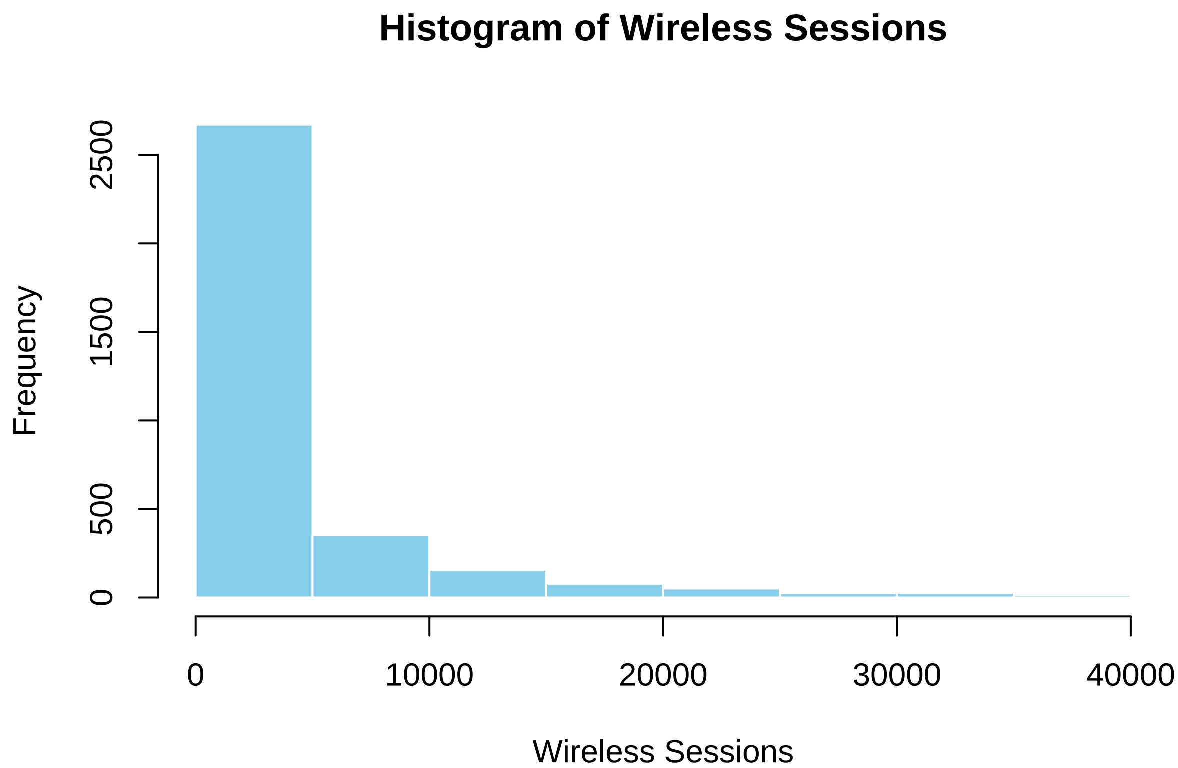

Histogram of Central Library
Frequency
0
500
1500
2500
3500
0.0
0.2
0.4
0.6
0.8
1.0
Central Library

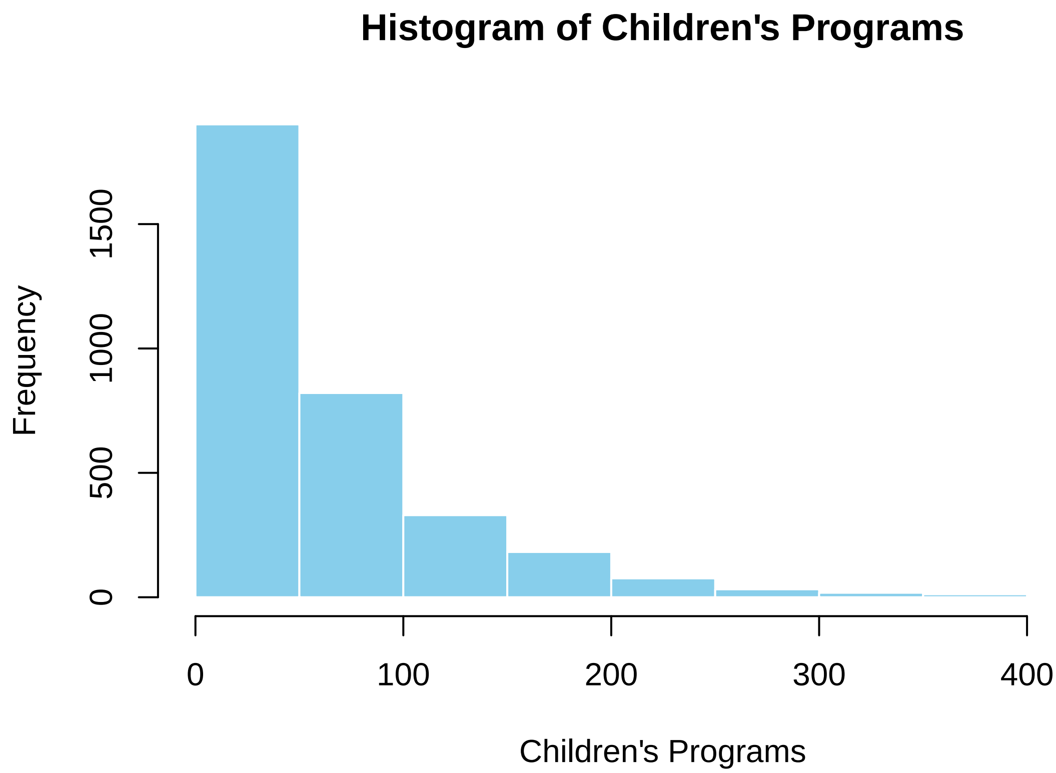

Histogram of Branch Library
Frequency
0
500
1500
2500
0
1
2
3
4
5
Branch Library

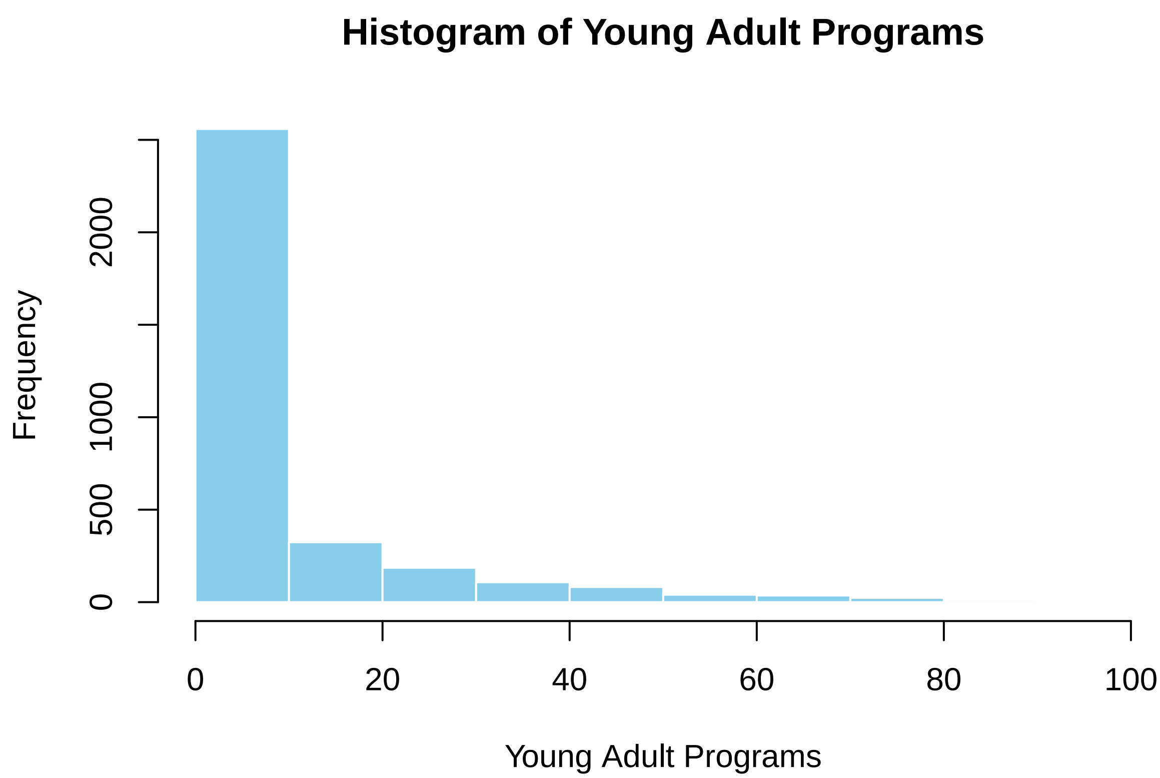

Histogram of Bookmobiles
Frequency
0
500
1500
2500
0.0
0.5
1.0
1.5
2.0
Bookmobiles

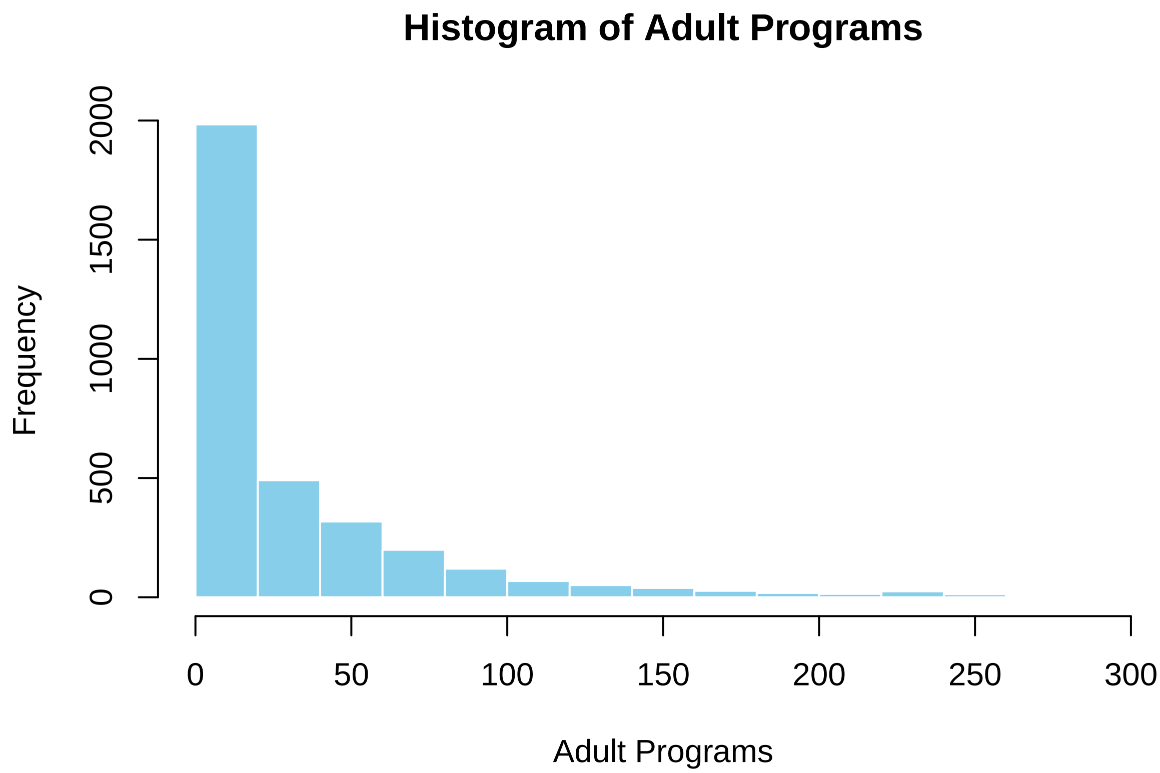

Histogram of Internet Computers
Frequency
0
500
1000
1500
0
10
20
30
40
50
Internet Computers

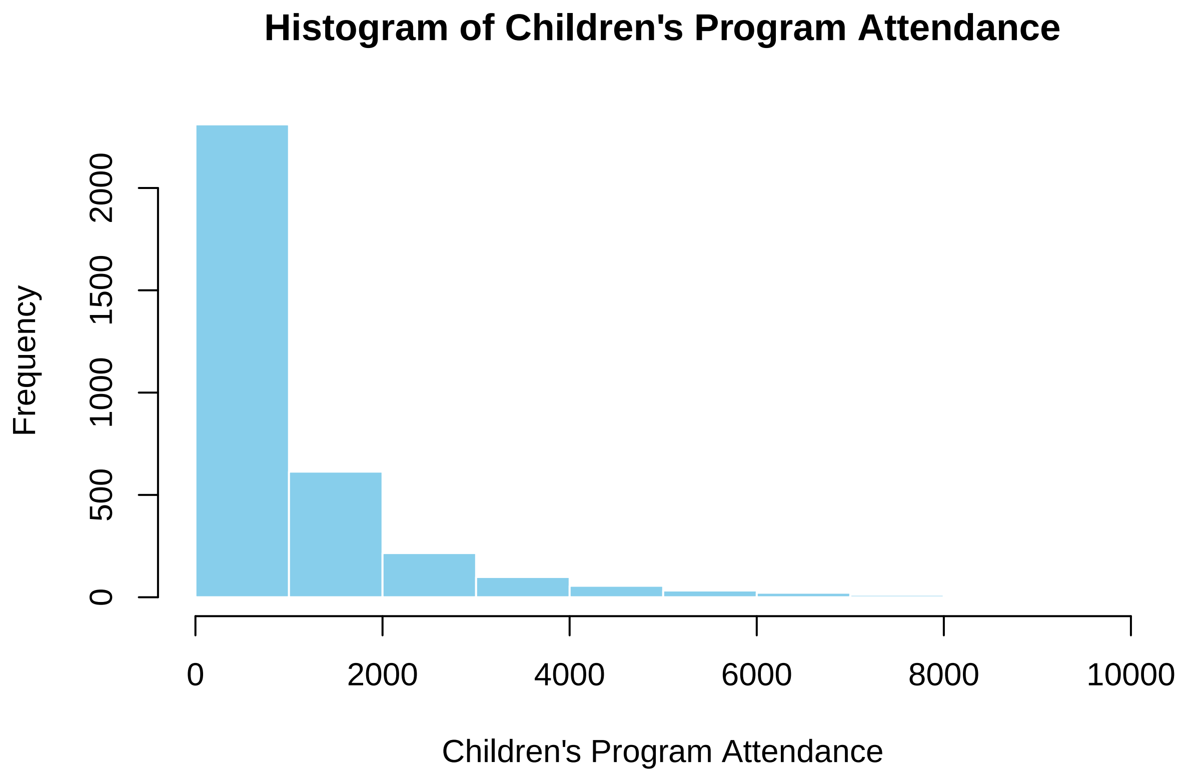

Histogram of Computer Uses
Frequency
0
500
1000
1500
2000
0
2000
4000
6000
8000
10000
12000
14000
Computer Uses

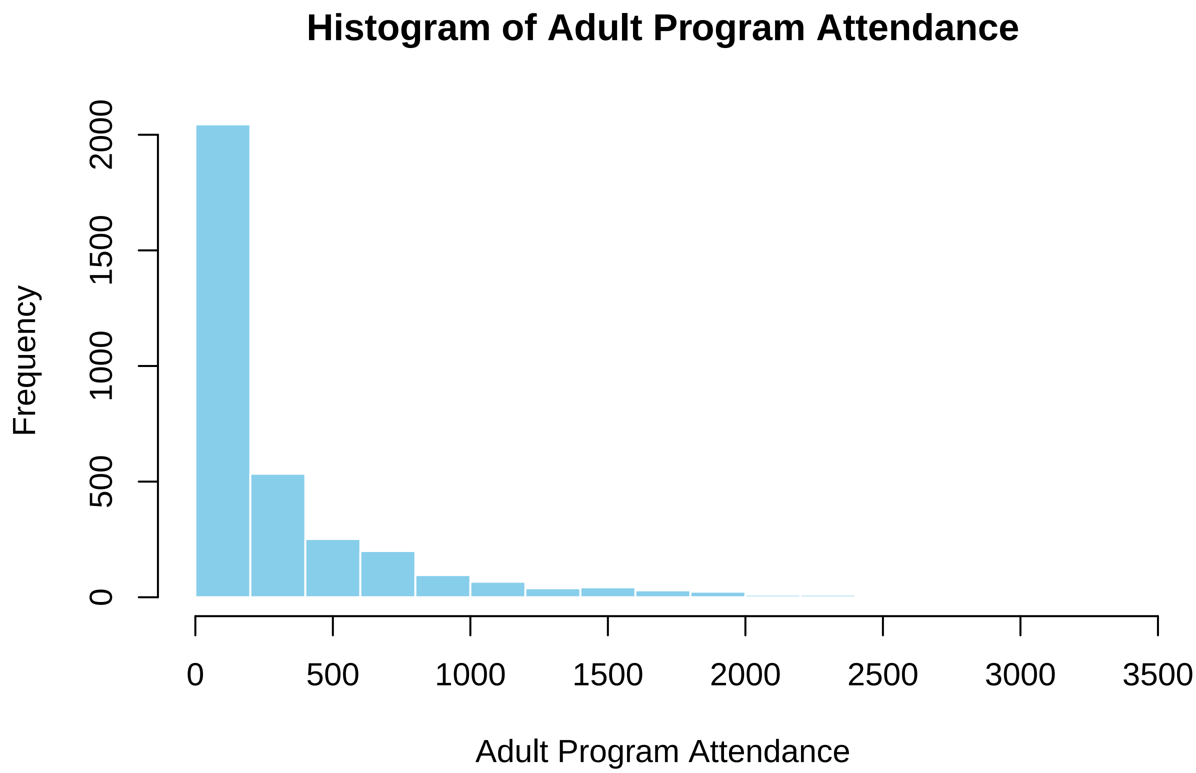

Histogram of Wireless Sessions
Frequency
0
500
1500
2500
0
10000
20000
30000
40000
Wireless Sessions


Histogram of Children's Programs
Frequency
0
500
1000
1500
0
100
200
300
400
Children's Programs

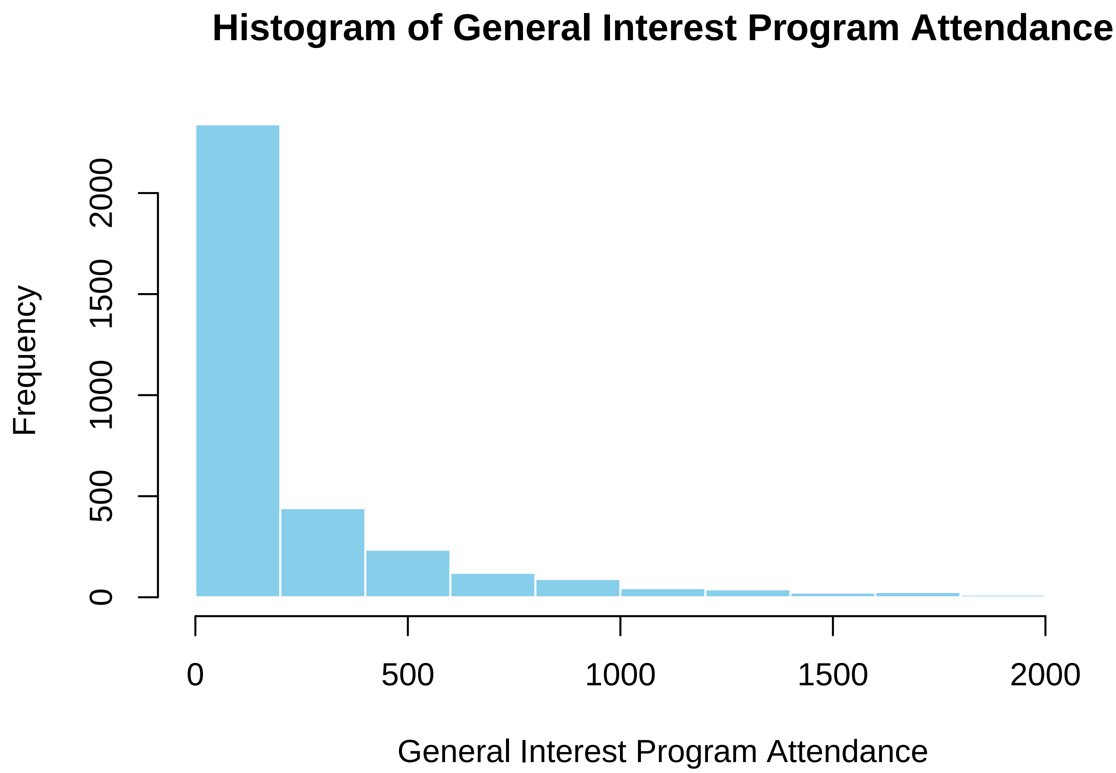

Histogram of Young Adult Programs
Frequency
0
500
1000
2000
0
20
40
60
80
100
Young Adult Programs

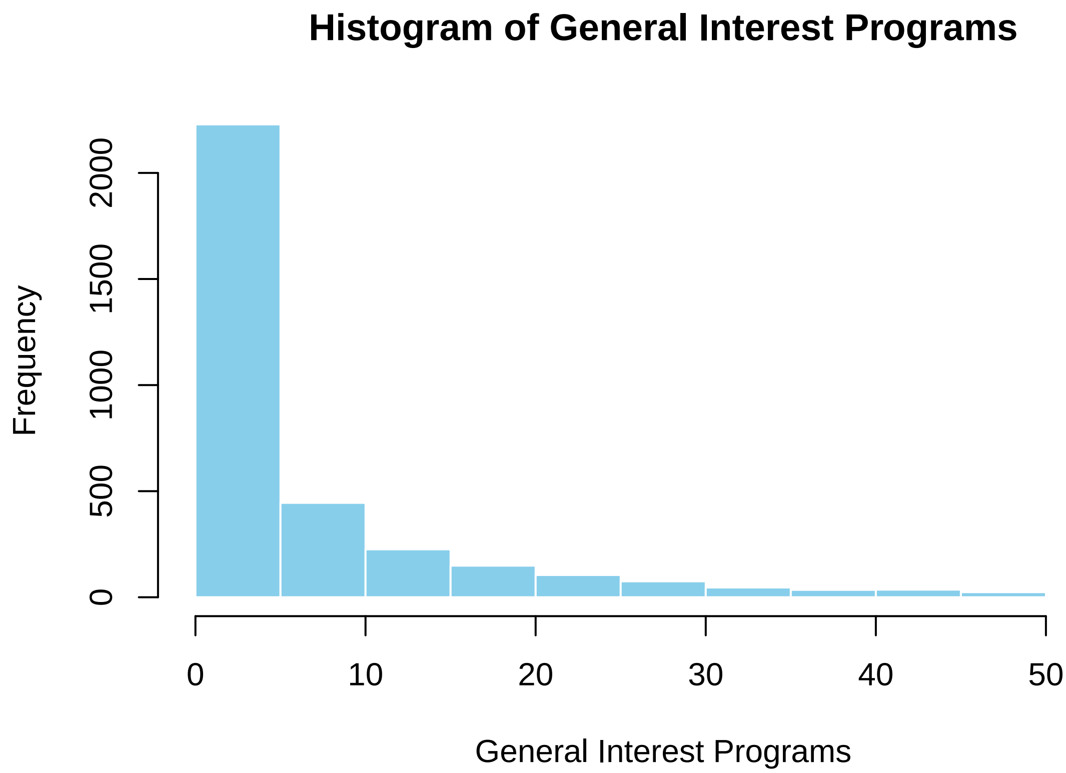

Histogram of Adult Programs
Frequency
0
500
1000
1500
2000
0
50
100
150
200
250
300
Adult Programs

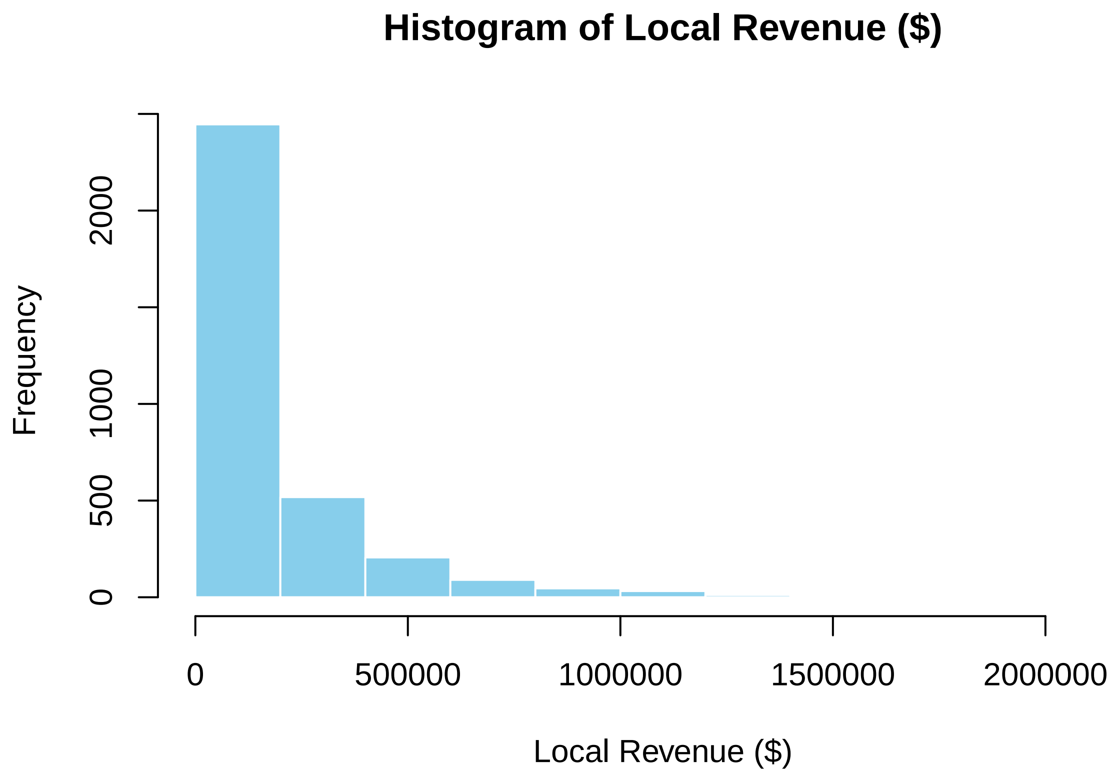

Histogram of Children's Program Attendance
Frequency
0
500
1000
1500
2000
0
2000
4000
6000
8000
10000
Children's Program Attendance

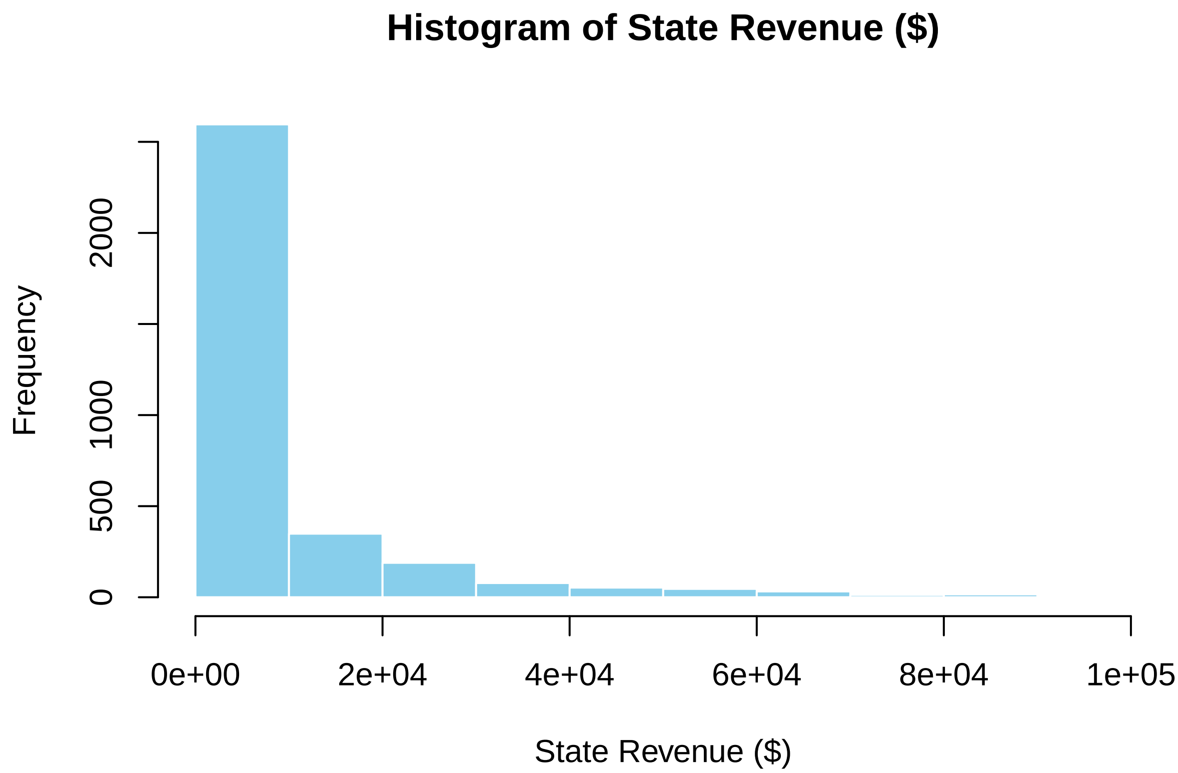


Histogram of Adult Program Attendance
Frequency
0
500
1000
1500
2000
0
500
1000
1500
2000
2500
3000
3500
Adult Program Attendance

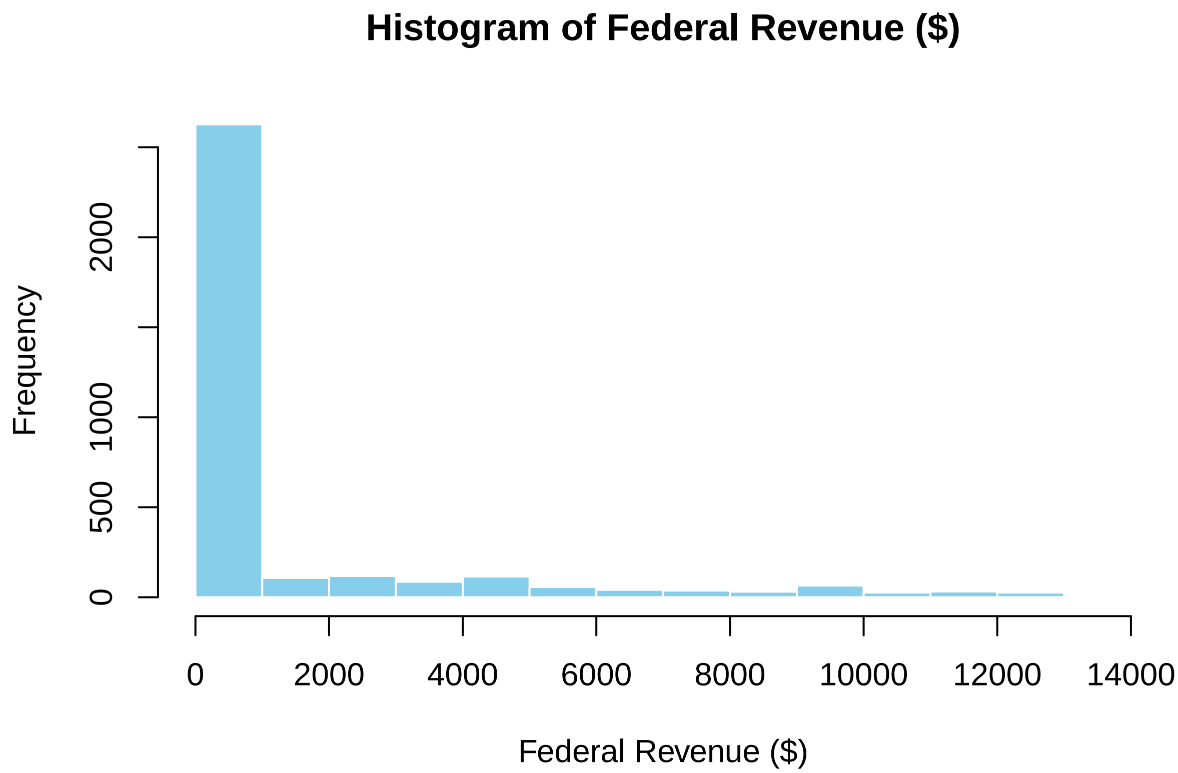

Histogram of General Interest Program Attendance
Frequency
0
500
1000
1500
2000
0
500
1000
1500
2000
General Interest Program Attendance


Histogram of General Interest Programs
Frequency
0
500
1000
1500
2000
0
10
20
30
40
50
General Interest Programs

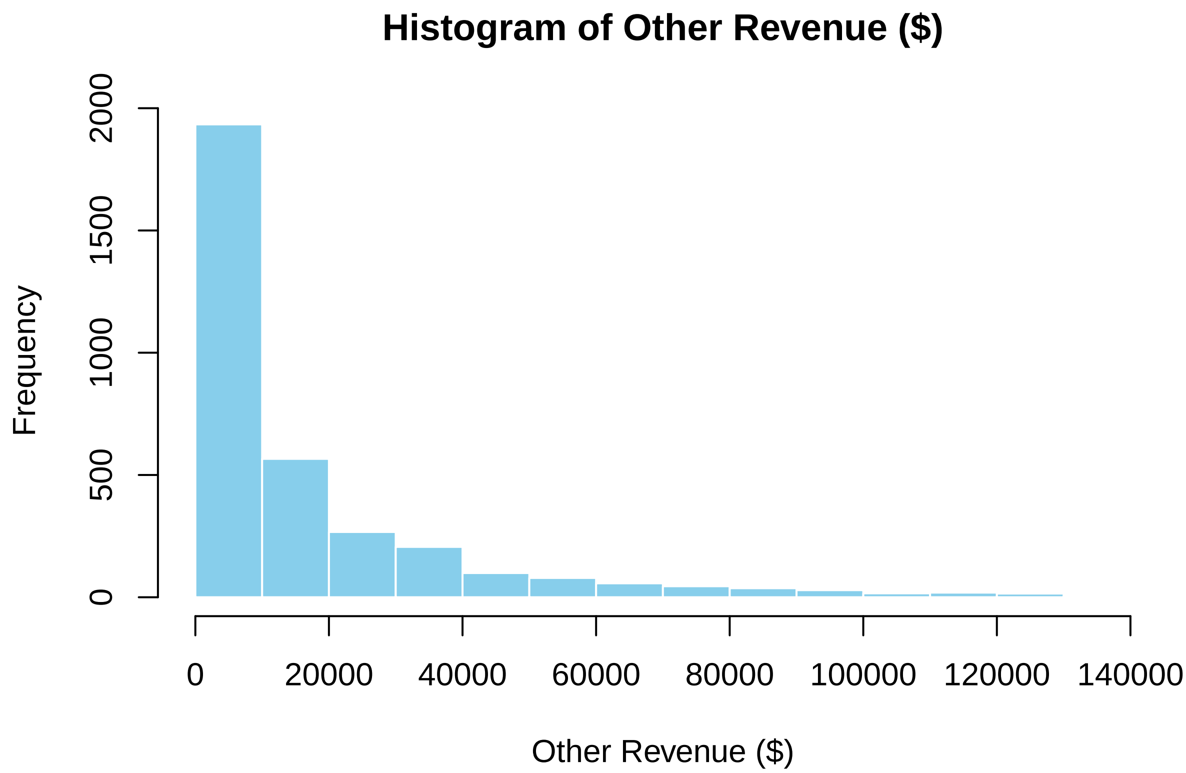

Histogram of Local Revenue ($)
Frequency
0
500
1000
2000
0
500000
1000000
1500000
2000000
Local Revenue ($)

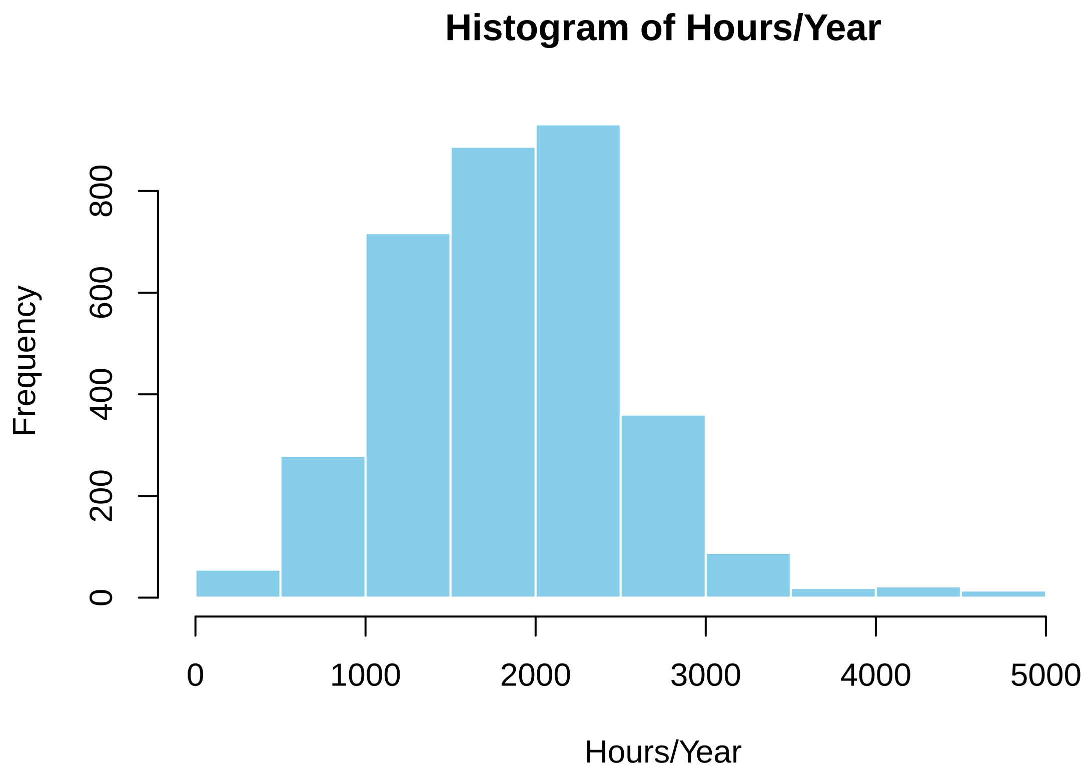


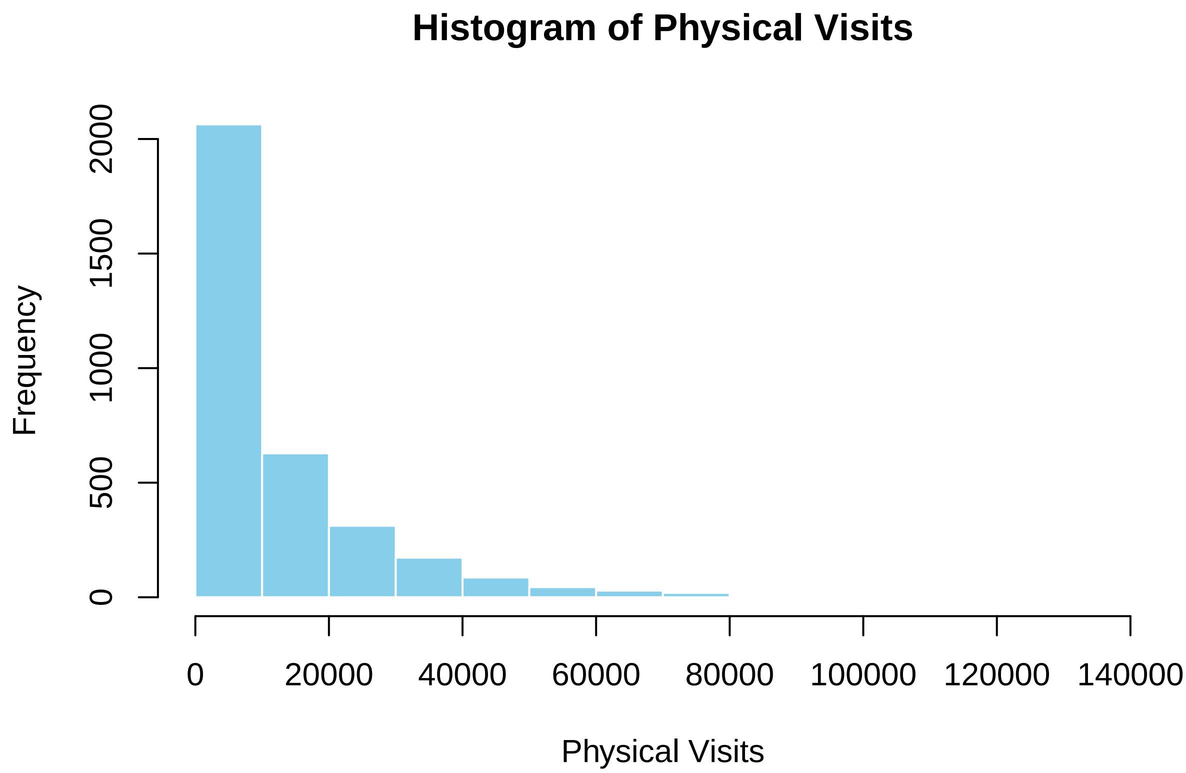
Histogram of State Revenue ($)
Frequency
0
500
1000
2000
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
State Revenue ($)

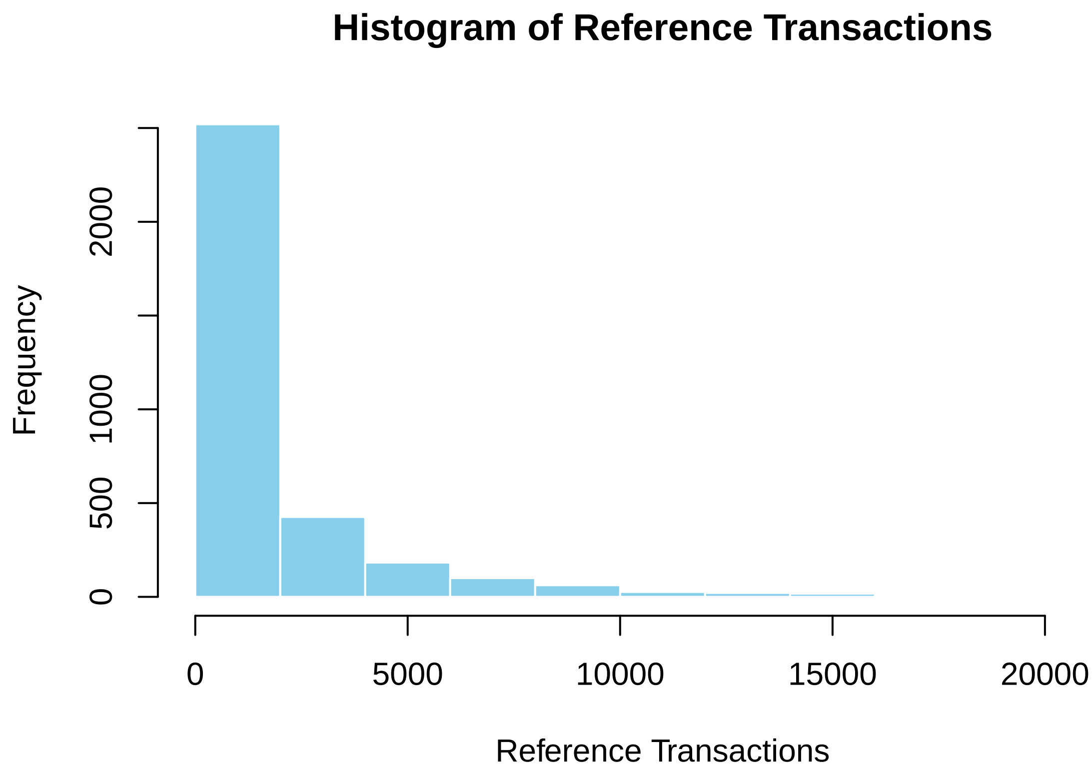


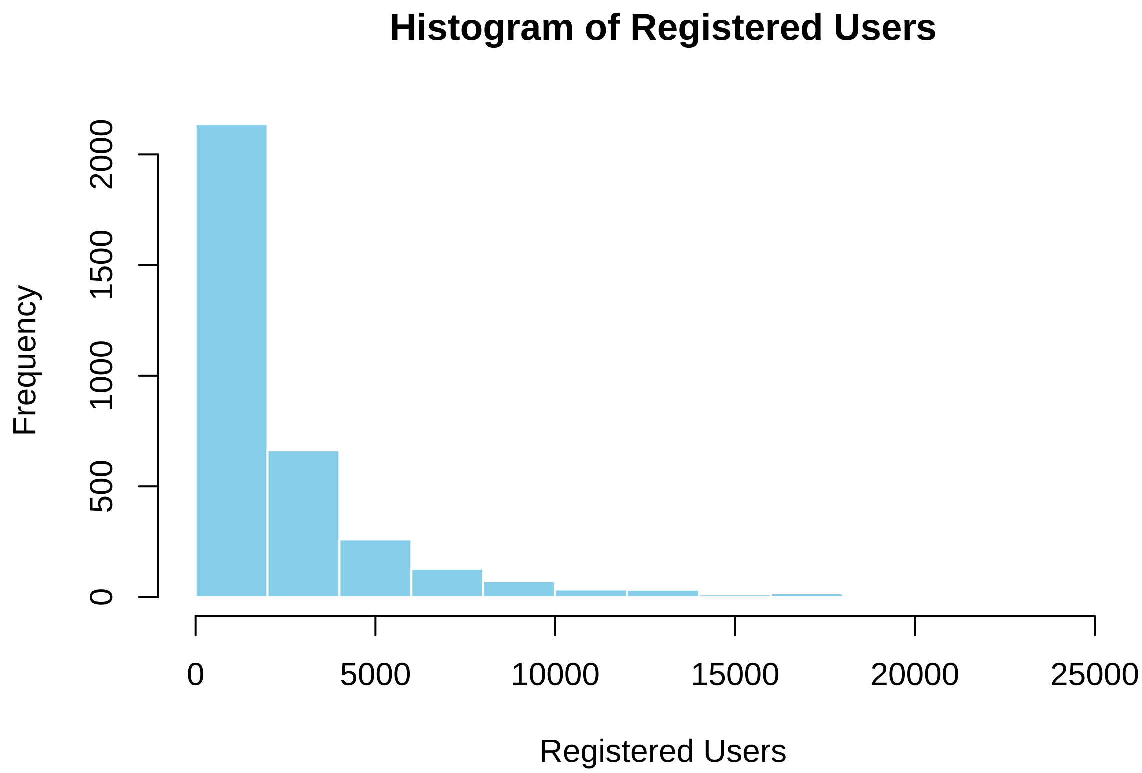
Histogram of Federal Revenue ($)
Frequency
0
500
1000
2000
0
2000
4000
6000
8000
10000
12000
14000
Federal Revenue ($)

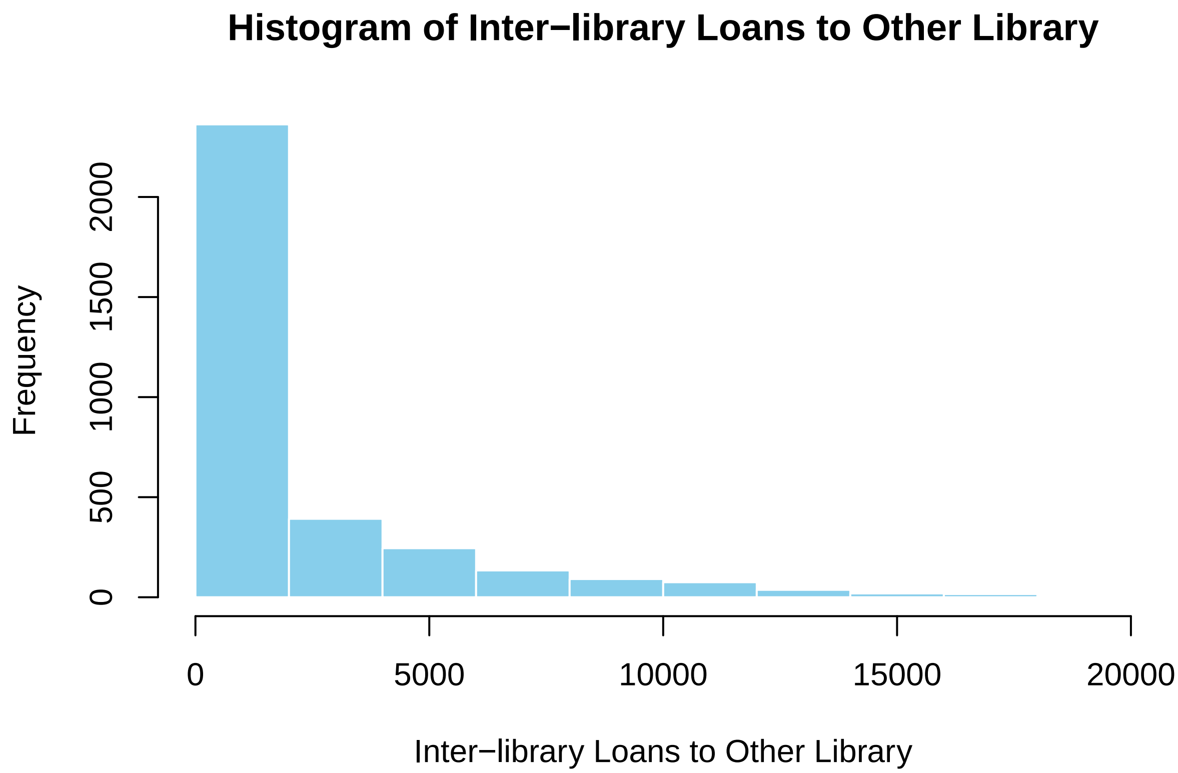

Histogram of Other Revenue ($)
Frequency
0
500
1000
1500
2000
0
20000
40000
60000
80000
100000
120000
140000
Other Revenue ($)

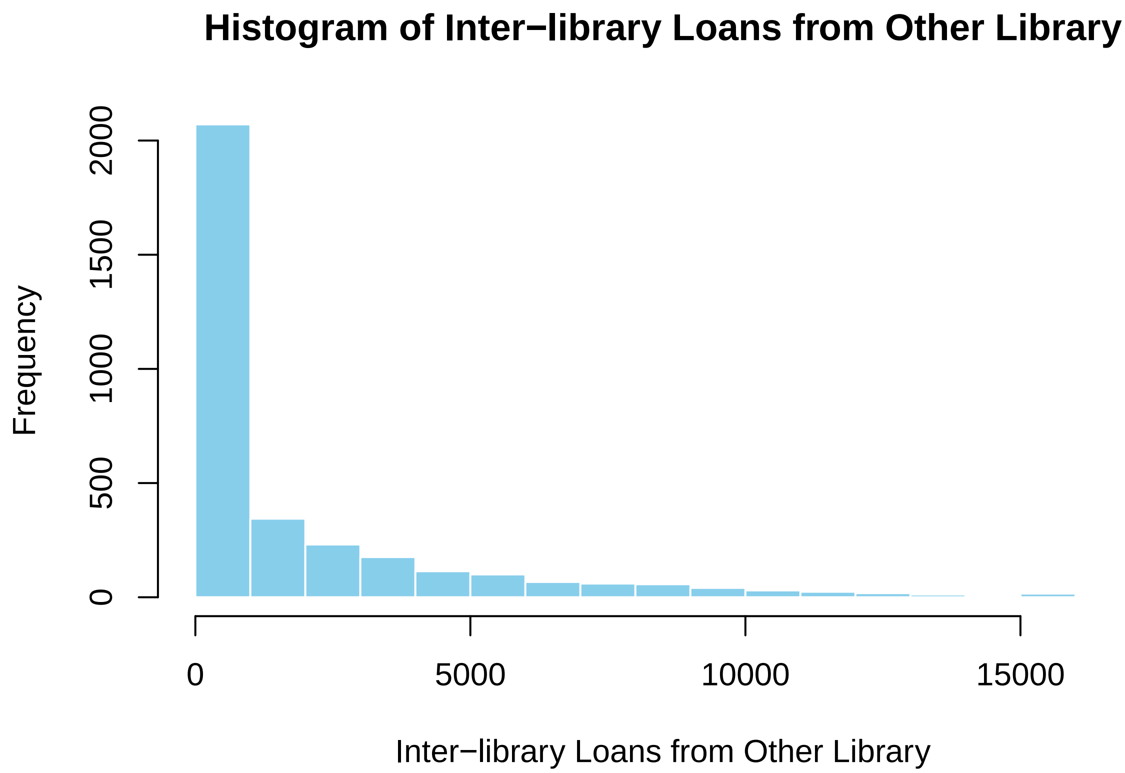

Histogram of Hours/Year
Frequency
0
200
400
600
800
0
1000
2000
3000
4000
5000
Hours/Year


Histogram of Physical Visits
Frequency
0
500
1000
1500
2000
0
20000
40000
60000
80000
100000
120000
140000
Physical Visits


Histogram of Reference Transactions
Frequency
0
500
1000
2000
0
5000
10000
15000
20000
Reference Transactions


Histogram of Registered Users
Frequency
0
500
1000
1500
2000
0
5000
10000
15000
20000
25000
Registered Users


Histogram of Inter−library Loans to Other Library
Frequency
0
500
1000
1500
2000
0
5000
10000
15000
20000
Inter−library Loans to Other Library

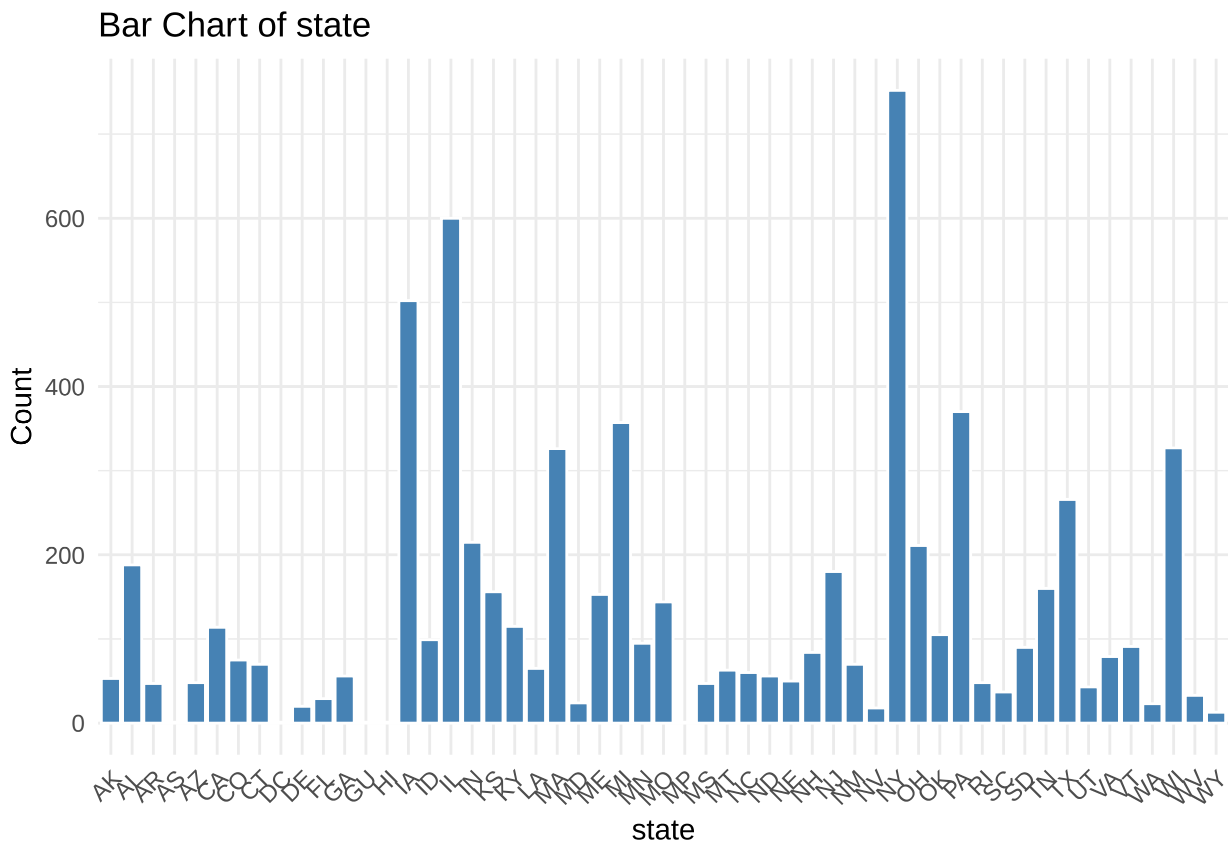


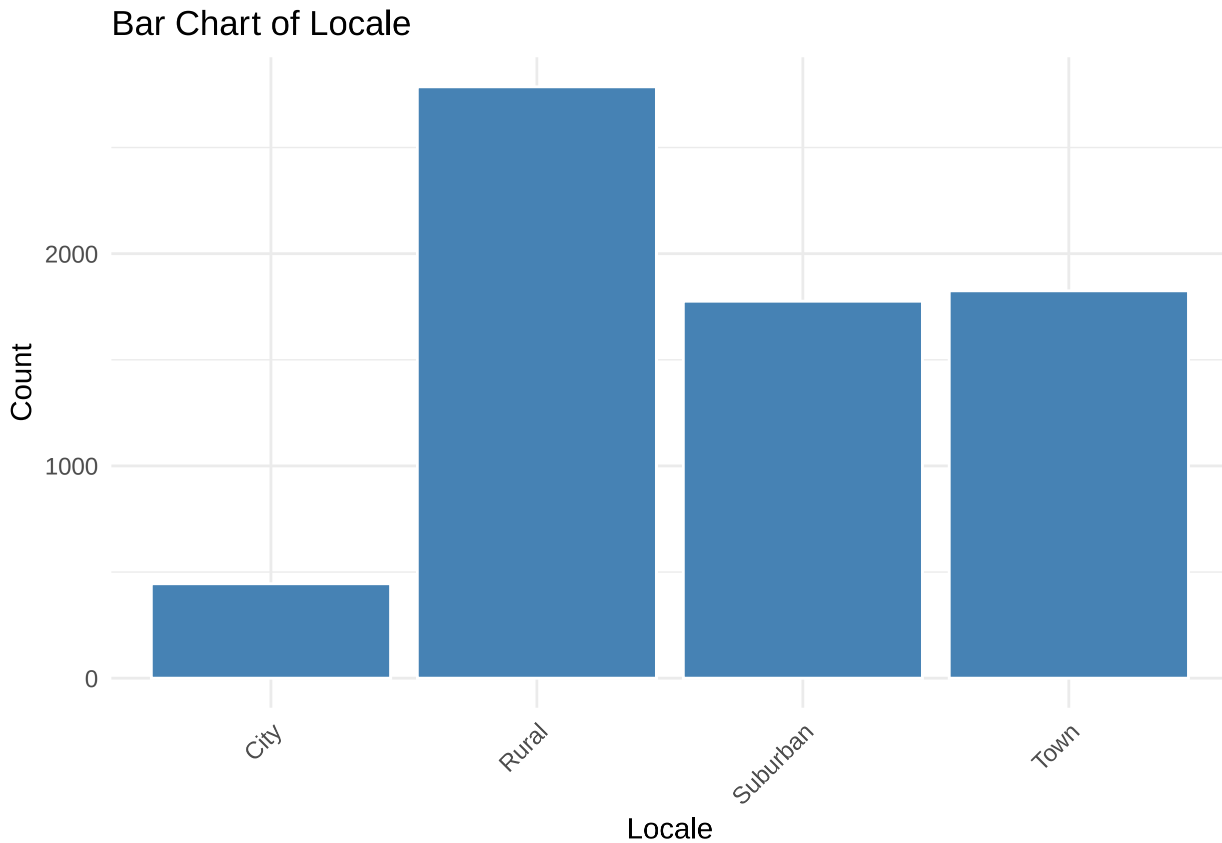
With outliers removed, the data is more reasonable, but skewed to the right. With this being said, we should consider Box-Cox transformations when making the models.

```
categorical_data <- libraries %>% select(!where(is.numeric))

# Loop through each categorical column and plot
for (col in names(categorical_data)) {
  ggplot(categorical_data, aes_string(x = col)) +
    geom_bar(fill = "steelblue", color = "white") +
    theme_minimal() +
    labs(
      title = paste("Bar Chart of", col),
      x = col,
      y = "Count"
    ) +
    theme(
      axis.text.x = element_text(angle = 45, hjust = 1)
    ) -> p

  print(p)
}
```

```
## Warning: 'aes_string()' was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with 'aes()'.
## i See also 'vignette("ggplot2-in-packages")' for more information.
## This warning is displayed once every 8 hours.
```

```
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```





With outliers removed, we see an overrepresentation of NY, IL, and HI, which means we should be careful making generalizations about the data to the entire nation. The locale is not as surprising, but we should be aware of the fact that the city is less represented than other types of locale.

Because this is a regression task, we should be considering Ridge and Lasso multi variable regression models to minimize the effects of the high collinearity.