Econ 104 Project 1

Anshika Khandelwal, Ishika Agrawal, Meghna Nair, Rajasvi Singh

2025-04-07

Contents

Question 1: Introducing Model & Question

This model explores what factors influence how often individuals visit Lake Somerville. Understanding these influences can help inform pricing strategies, investment in amenities, and policies that support local tourism and recreational planning, which can eventually lead to an increase in the economic prosperity of the area.

```
install.packages("Ecdat", repos = "https://cloud.r-project.org/")
```

```
##
## The downloaded binary packages are in
##  /var/folders/sh/y5krnstj3pn0wqd0wr_prc3m0000gn/T//RtmpcqjhCU/downloaded_packages
```

```
data(package = "Ecdat")
data("Somerville", package = "Ecdat")
head(Somerville)
```

```
##   visits quality ski income feeSom costCon costSom costHoust
## 1      0       0 yes      4     no   67.59  68.620    76.800
## 2      0       0  no      9     no   68.86  70.936    84.780
## 3      0       0 yes      5     no   58.12  59.465    72.110
## 4      0       0  no      2     no   15.79  13.750    23.680
## 5      0       0 yes      3     no   24.02  34.033    34.547
## 6      0       0 yes      5     no  129.46 137.377   137.850
```

```
summary (Somerville)
```

```
##      visits          quality       ski          income       feeSom
##  Min.   : 0.000   Min.   :0.000   no :417   Min.   :1.000   no :646
##  1st Qu.: 0.000   1st Qu.:0.000   yes:242   1st Qu.:3.000   yes: 13
##  Median : 0.000   Median :0.000             Median :3.000
##  Mean   : 2.244   Mean   :1.419             Mean   :3.853
##  3rd Qu.: 2.000   3rd Qu.:3.000             3rd Qu.:5.000
##  Max.   :88.000   Max.   :5.000             Max.   :9.000
##     costCon          costSom          costHoust
##  Min.   :  4.34   Min.   :  4.767   Min.   :  5.70
##  1st Qu.: 28.24   1st Qu.: 33.312   1st Qu.: 28.96
##  Median : 41.19   Median : 47.000   Median : 42.38
##  Mean   : 55.42   Mean   : 59.928   Mean   : 55.99
##  3rd Qu.: 69.67   3rd Qu.: 72.573   3rd Qu.: 68.56
##  Max.   :493.77   Max.   :491.547   Max.   :491.05
```

Question 2: Describing the Dataset

(a) Citations

Dataset: Somerville – Visits to Lake Somerville Source: Ecdat package in R, drawn from a 1980 cross-sectional survey conducted in the United States. Citation: Crooker, J. R., & Herriges, J. A. (2004). "Do fish bite when it's raining? Evidence from the demand for recreational fishing," Journal of Agricultural and Resource Economics. Accessed via the Ecdat R package.

(b) Summary

The dataset contains 659 individual-level observations from a 1980 cross-sectional study in the United States. It investigates factors influencing the annual number of visits individuals made to Lake Somerville, along with variables reflecting personal preferences, household income, and costs associated with visiting nearby recreational lakes.

(c) Descriptive analysis of variables

Variables: quality, income, costCon, costSom, costHoust

```
Somerville <- Ecdat::Somerville
str(Somerville)
```

```
## 'data.frame':    659 obs. of  8 variables:
##  $ visits   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ quality  : num  0 0 0 0 0 0 0 0 0 2 ...
##  $ ski      : Factor w/ 2 levels "no","yes": 2 1 2 1 2 2 1 2 1 1 ...
##  $ income   : num  4 9 5 2 3 5 1 5 2 3 ...
##  $ feeSom   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ costCon  : num  67.6 68.9 58.1 15.8 24 ...
##  $ costSom  : num  68.6 70.9 59.5 13.8 34 ...
##  $ costHoust: num  76.8 84.8 72.1 23.7 34.5 ...
```

```
summary(Somerville)
```

```
##      visits          quality        ski          income       feeSom
##  Min.   : 0.000   Min.   :0.000   no :417   Min.   :1.000   no :646
##  1st Qu.: 0.000   1st Qu.:0.000   yes:242   1st Qu.:3.000   yes: 13
##  Median : 0.000   Median :0.000             Median :3.000
##  Mean   : 2.244   Mean   :1.419             Mean   :3.853
##  3rd Qu.: 2.000   3rd Qu.:3.000             3rd Qu.:5.000
##  Max.   :88.000   Max.   :5.000             Max.   :9.000
##     costCon          costSom          costHoust
##  Min.   :  4.34   Min.   :  4.767   Min.   :  5.70
##  1st Qu.: 28.24   1st Qu.: 33.312   1st Qu.: 28.96
##  Median : 41.19   Median : 47.000   Median : 42.38
##  Mean   : 55.42   Mean   : 59.928   Mean   : 55.99
##  3rd Qu.: 69.67   3rd Qu.: 72.573   3rd Qu.: 68.56
##  Max.   :493.77   Max.   :491.547   Max.   :491.05
```

```
install.packages("ggplot2", repos = "https://cloud.r-project.org/")
```

```
##
## The downloaded binary packages are in
##  /var/folders/sh/y5krnstj3pn0wqd0wr_prc3m0000gn/T//RtmpcqjhCU/downloaded_packages
```

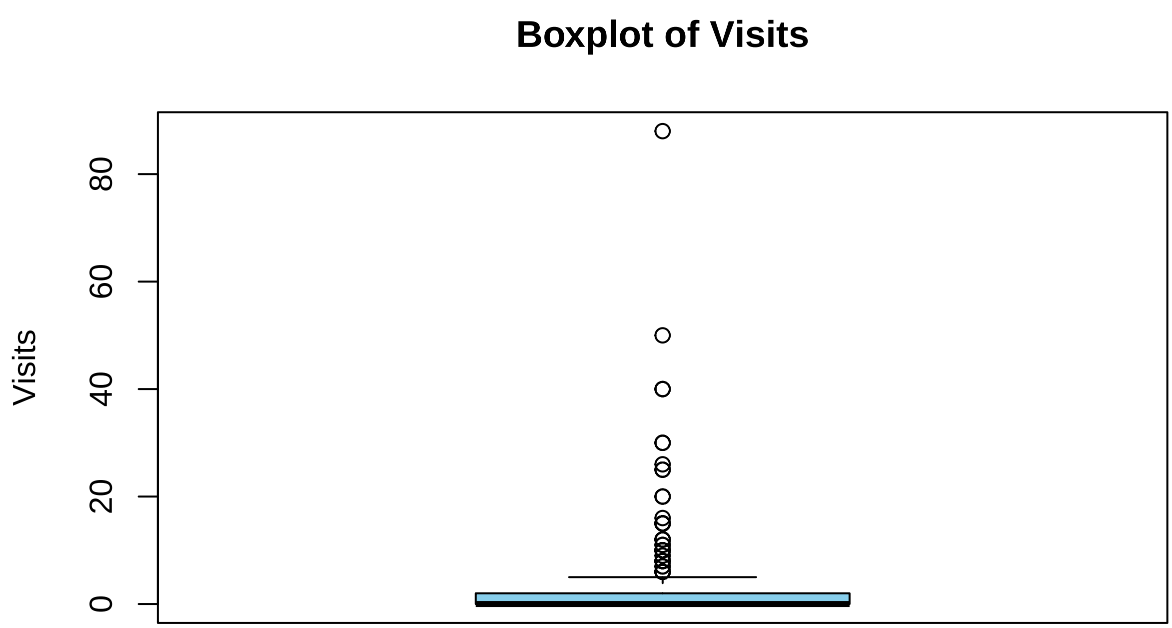
```r
install.packages("corrplot", repos = "https://cloud.r-project.org/")
```

```
##
## The downloaded binary packages are in
##  /var/folders/sh/y5krnstj3pn0wqd0wr_prc3m0000gn/T//RtmpcqjhCU/downloaded_packages
```

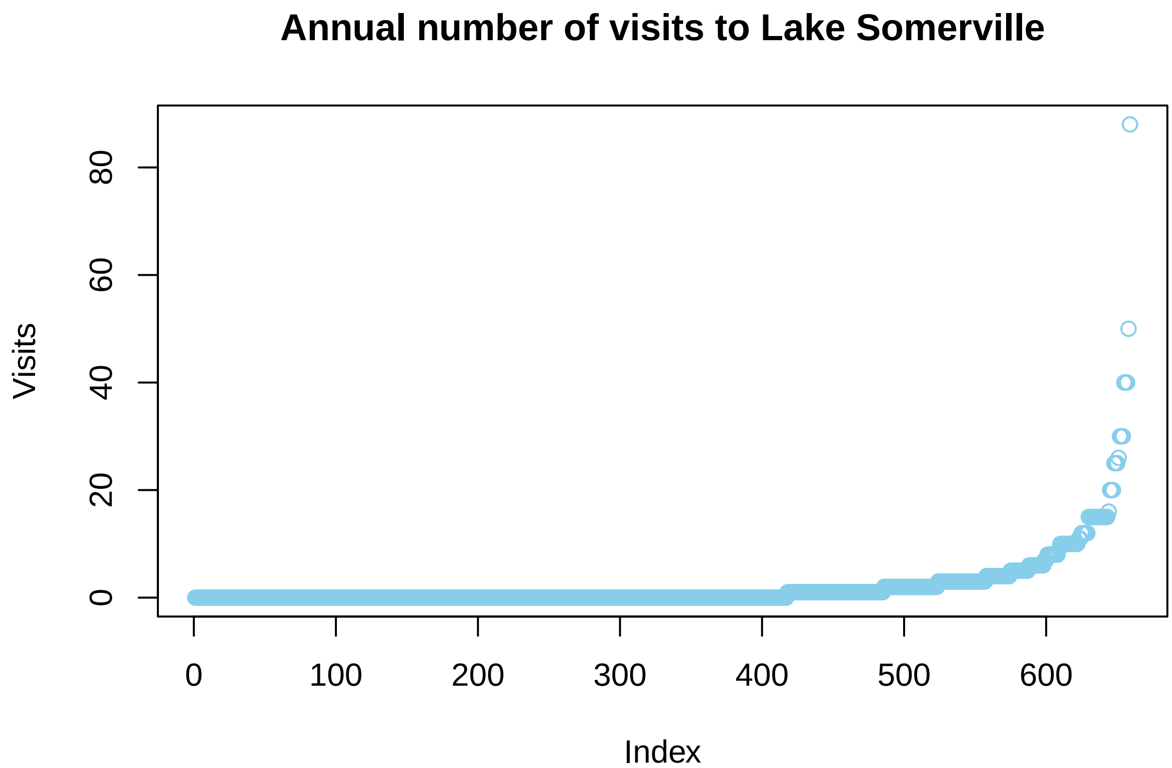
```r
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

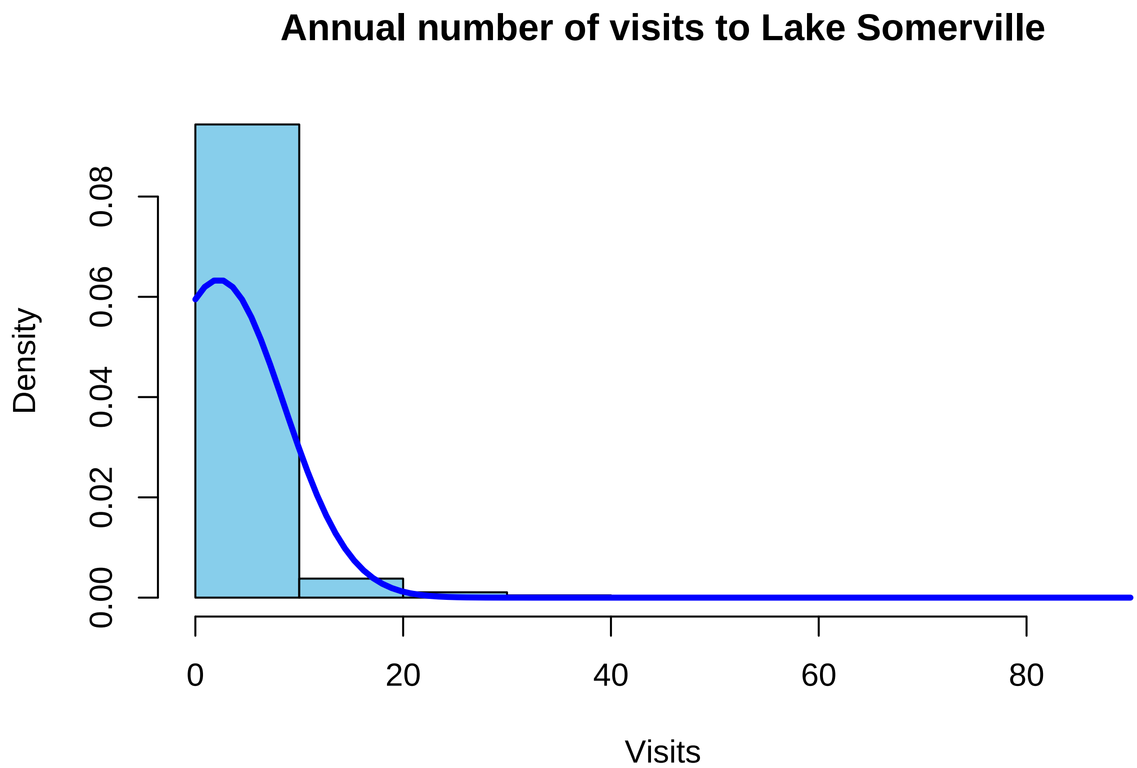
```r
boxplot(Somerville$visits, main = "Boxplot of Visits", col = "skyblue", ylab = "Visits")
```




```r
plot(Somerville$visits,col = "skyblue",
     main = "Annual number of visits to Lake Somerville",
     ylab = "Visits")
```




```r
hist(Somerville$visits, main = "Annual number of visits to Lake Somerville",
     xlab = "Visits", col = "skyblue",
     probability = TRUE)
curve(dnorm(x, mean = mean(Somerville$visits, na.rm = TRUE),
            sd = sd(Somerville$visits, na.rm = TRUE)),
      col = "blue", lwd = 3, add = TRUE)
```




```r
summary(Somerville$visits)
```

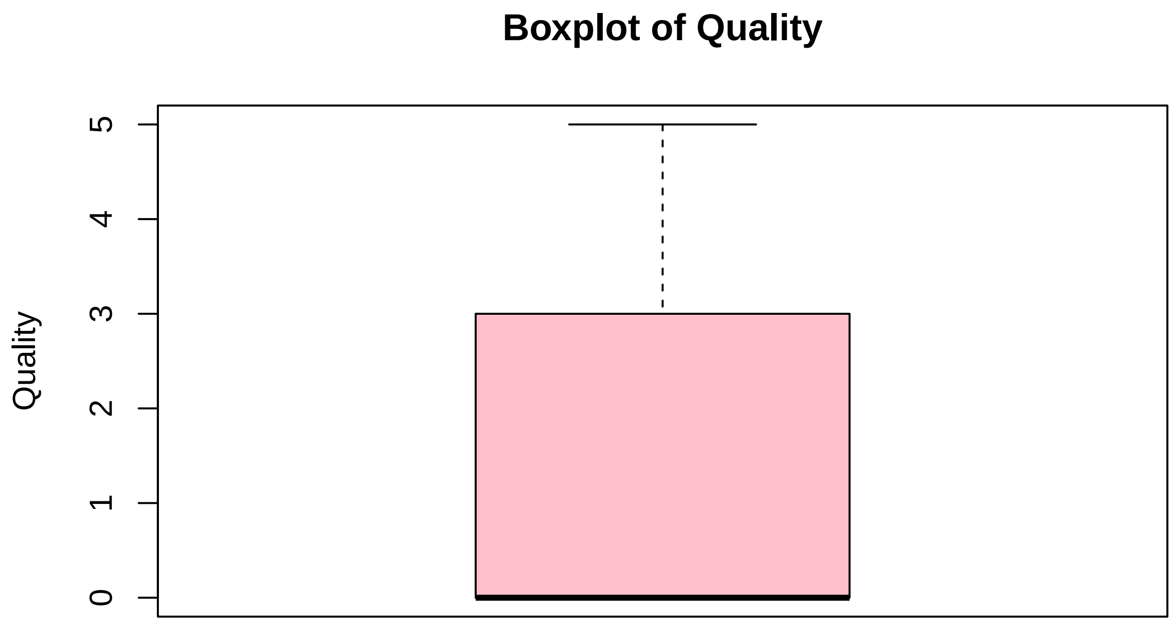
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   2.244   2.000  88.000
```

The distribution of lake quality rankings score is heavily right-skewed, with the majority of responses clustered at 0. This suggests that not many people make annual visits at Lake Somerville. The central tendency is low. There's a wide spread overall, but the concentration near 0 points to generally negative perceptions of the lake Somerville's recreational potential.

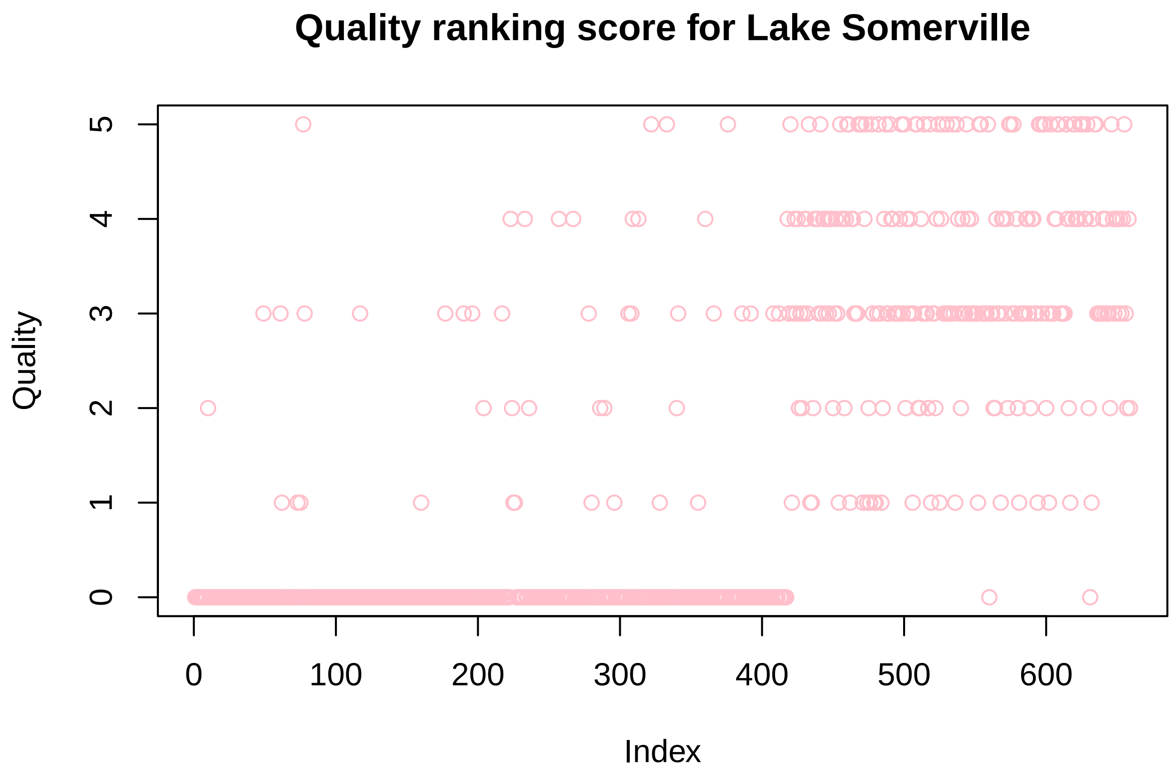
The boxplot shows that most visit counts are clustered at the lower end, with a large number of outliers extending above the upper whisker, indicating a right-skewed distribution. The median is low, suggesting that the majority of individuals have few visits annually.

The scatterplot displays a heavily right-skewed distribution with most data points near zero and a few extreme high values. This suggests that while most people visit infrequently, a small number of individuals visit Lake Somerville very frequently.

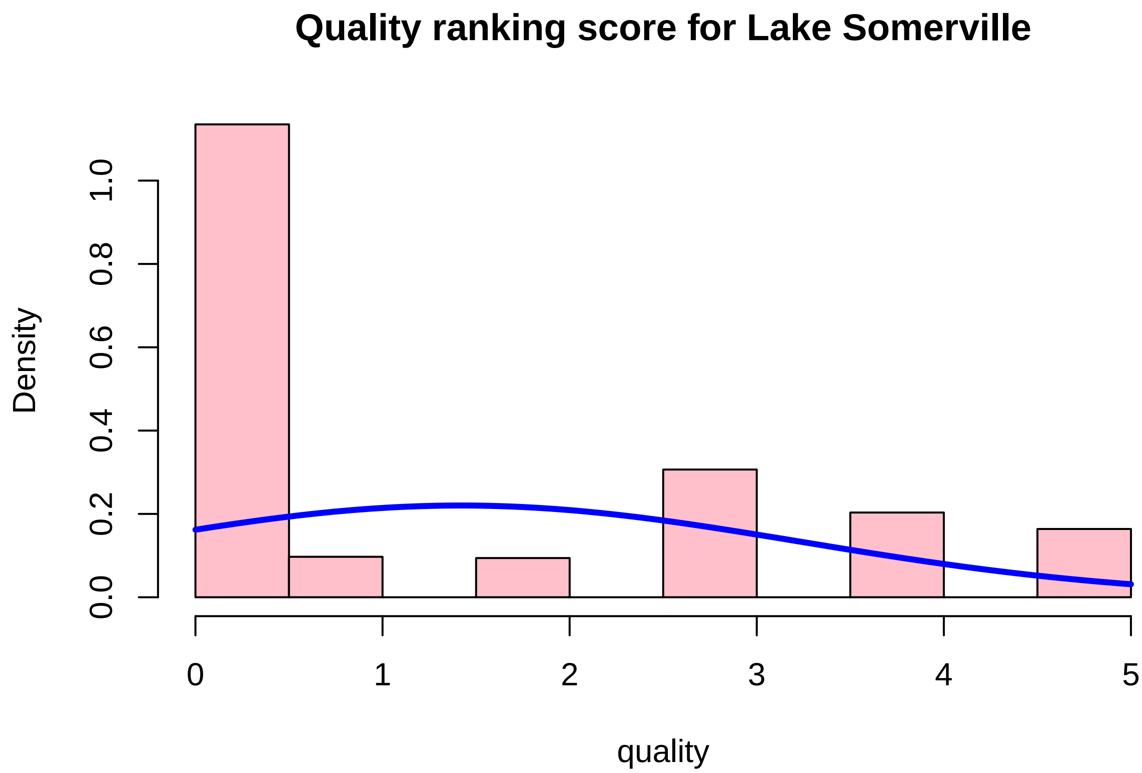
```r
library(corrplot)
boxplot(Somerville$quality, main = "Boxplot of Quality",
        col = "pink", ylab = "Quality")
```




```r
plot(Somerville$quality,col = "pink",
     main = "Quality ranking score for Lake Somerville",
     ylab = "Quality")
```




```r
hist(Somerville$quality, main = "Quality ranking score for Lake Somerville",
     col = "pink", probability = TRUE, xlab = "quality")
curve(dnorm(x, mean = mean(Somerville$quality, na.rm = TRUE),
            sd = sd(Somerville$quality, na.rm = TRUE)), col = "blue", lwd = 3, add = TRUE)
```




```r
summary(Somerville$quality)
```

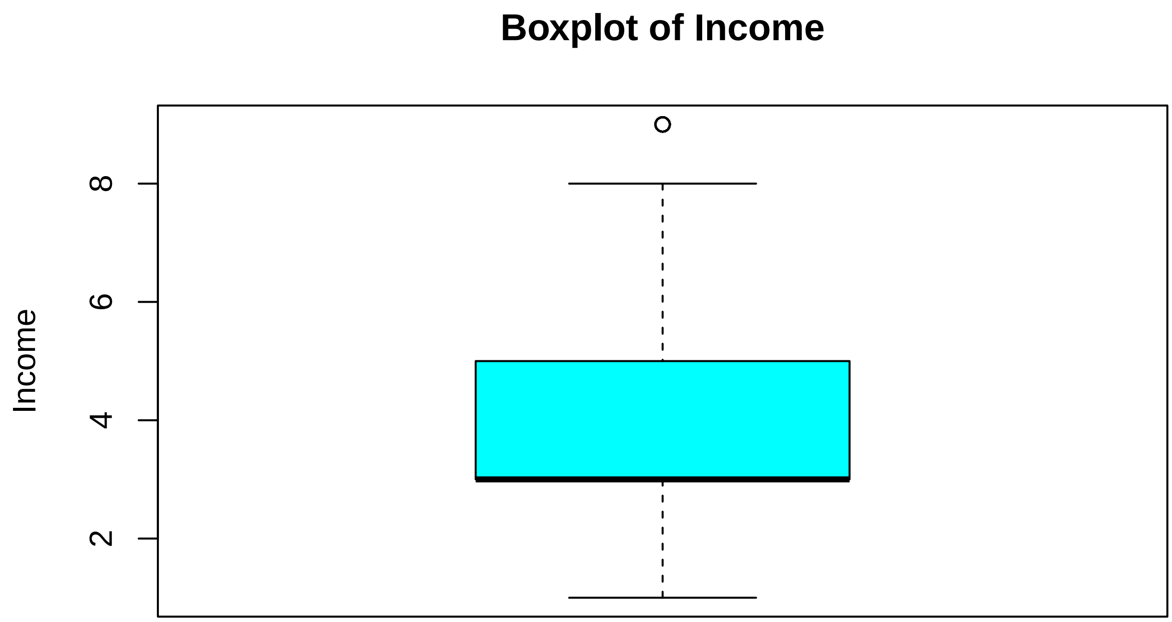
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.419   3.000   5.000
```

The distribution of lake quality rankings score is heavily right-skewed, with the majority of responses clustered at 0. This suggests that many individuals rated the lake as being of poor quality. The central tendency is low, and while a few respondents gave higher scores, those are much less frequent. There's a wide spread overall, but the concentration near 0 points to generally negative perceptions of the lake Someerville's condition.

The boxplot shows a highly skewed distribution with a median around 3 and a large cluster of values at 0, suggesting a significant portion of low-quality scores with a wide spread up to the maximum value of 5.

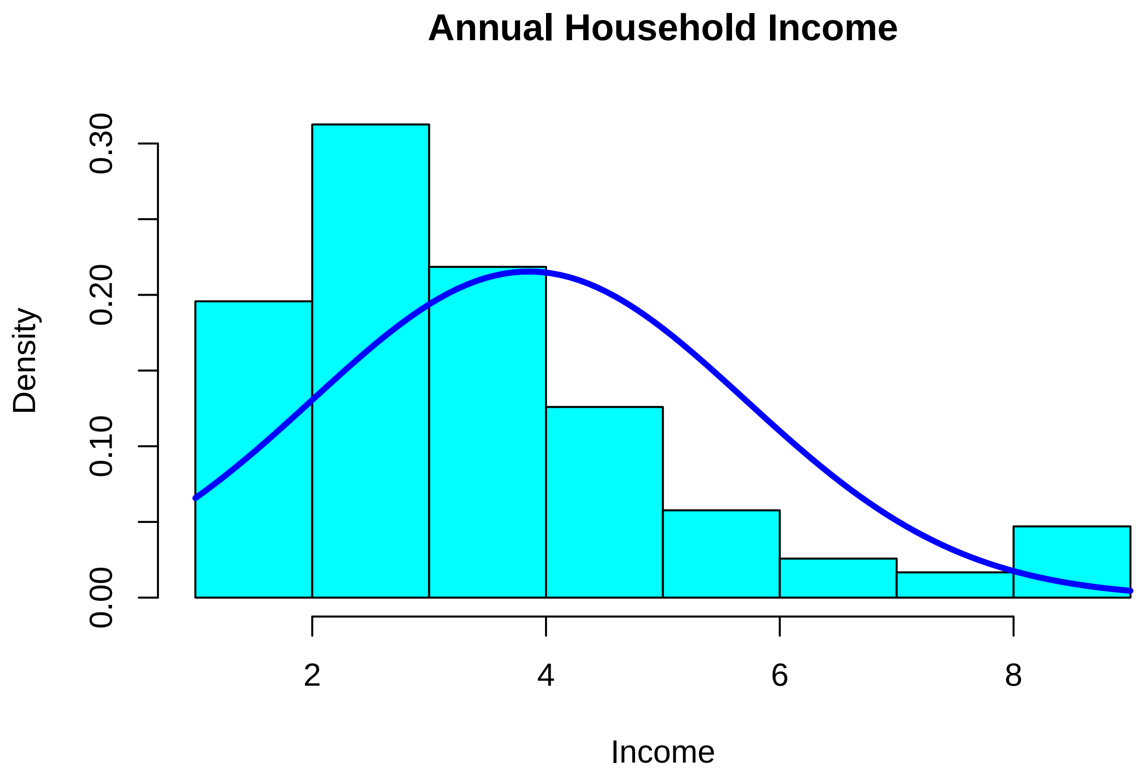
The scatterplot shows that a large number of observations are concentrated at Quality score 0, while scores from 1 to 5 appear less frequently but are more evenly spread after index 400, indicating a potential change in data pattern or collection method.

```r
boxplot(Somerville$income, main = "Boxplot of Income",
        col = "cyan", ylab = "Income")
```




```r
plot(Somerville$income,col = "cyan",
     main = "Annual Household Income", ylab = "Income")
```


```r
hist(Somerville$income, main = "Annual Household Income",
     xlab = "Income", col = "cyan", probability = TRUE)
curve(dnorm(x, mean = mean(Somerville$income, na.rm = TRUE),
            sd = sd(Somerville$income, na.rm = TRUE)), col = "blue", lwd = 3, add = TRUE)
```




```r
summary(Somerville$income)
```

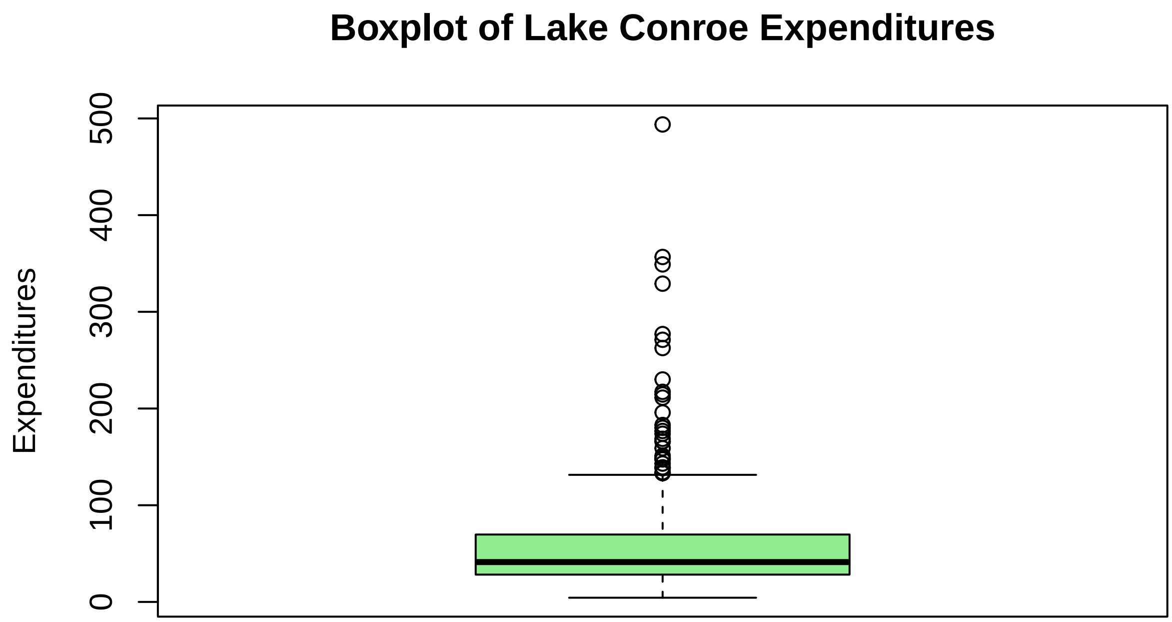
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.853   5.000   9.000
```

The distribution of income is moderately right-skewed, with most observations falling on the lower end of the scale. The central tendency appears to be between 2 and 4, meaning most respondents fall in the lower to middle income brackets. There's some noticeable spread, especially with a few higher-income outliers, which increases the overall dispersion. This suggests income levels vary quite a bit across the sample, though the majority are still clustered toward the bottom.

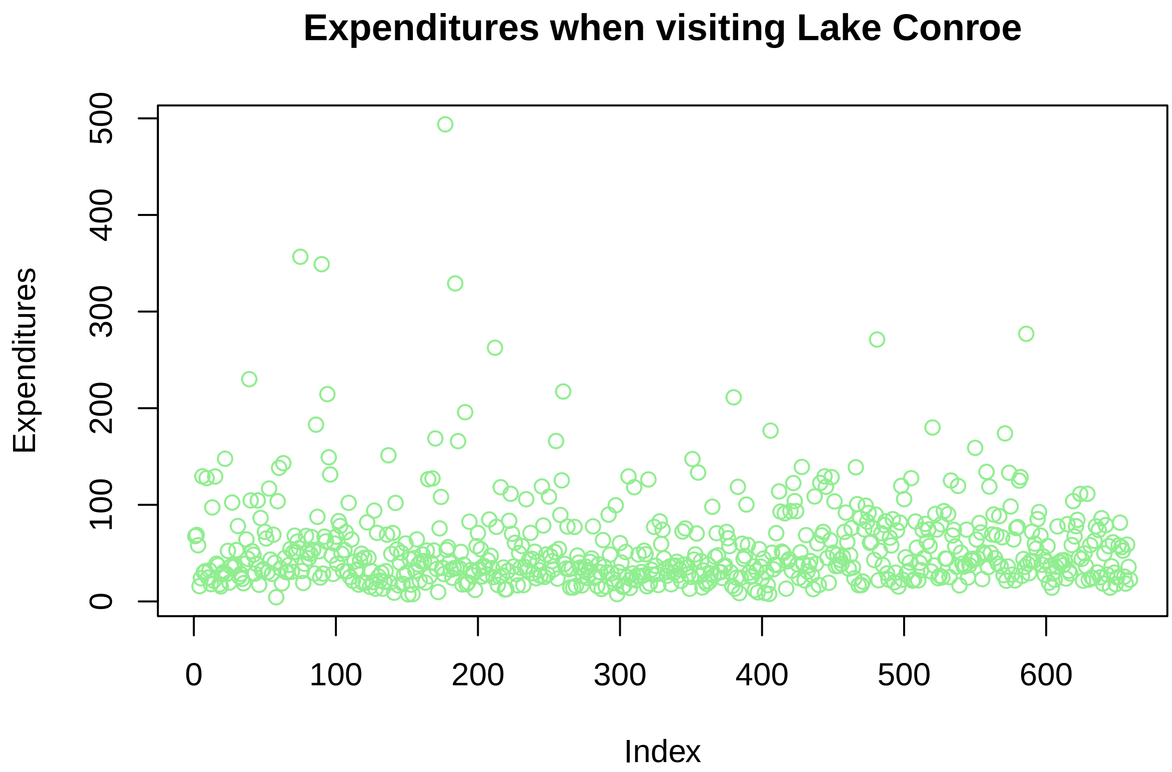
The boxplot of income shows a fairly symmetric distribution with a median around 4.5. Most income values lie between 3 and 6, with one noticeable outlier above 8, indicating a relatively high income compared to the rest.

The scatter plot titled "Annual Household Income" shows individual income data points plotted against their index values. Most data points cluster between income levels 2 and 5, indicating a concentration of households in that income range.

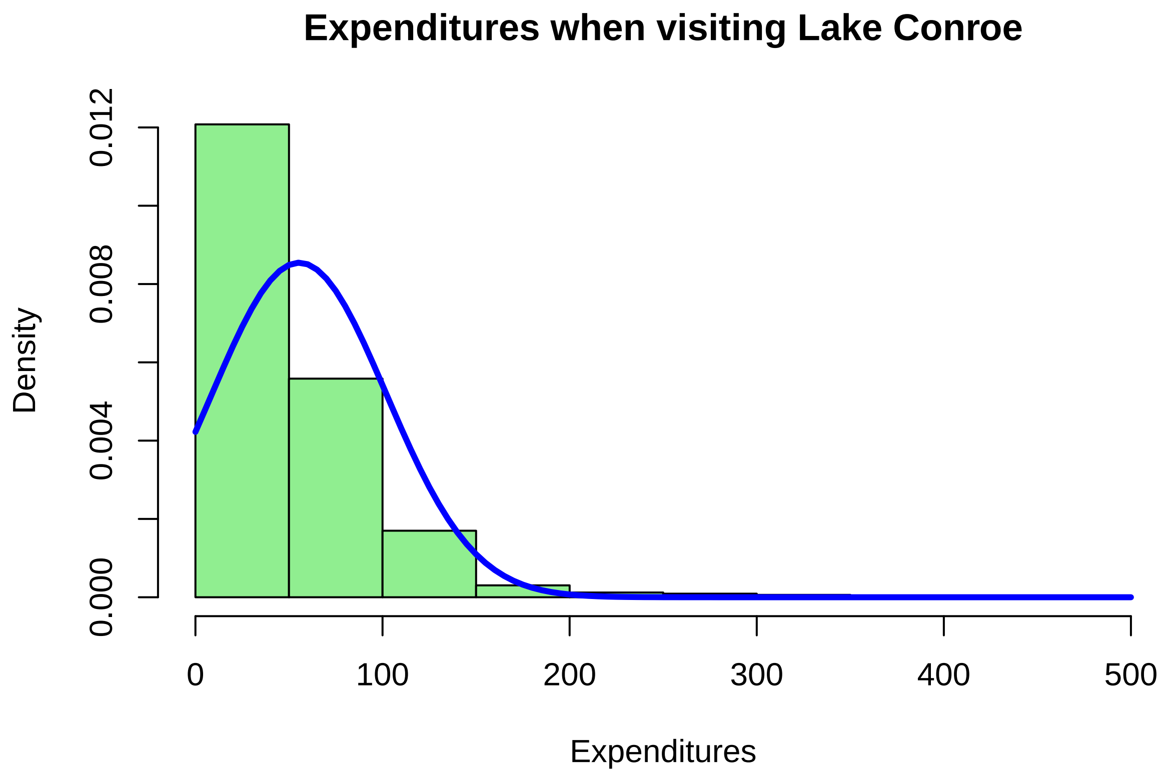
```r
boxplot(Somerville$costCon, main = "Boxplot of Lake Conroe Expenditures",
        ylab = "Expenditures", col = "lightgreen")
```




```
plot(Somerville$costCon,col = "lightgreen",
     main = "Expenditures when visiting Lake Conroe", ylab = "Expenditures")
```




```r
hist(Somerville$costCon, main = "Expenditures when visiting Lake Conroe",
     xlab = "Expenditures", col = "lightgreen", probability = TRUE)
curve(dnorm(x, mean = mean(Somerville$costCon, na.rm = TRUE),
            sd = sd(Somerville$costCon, na.rm = TRUE)), col = "blue", lwd = 3, add = TRUE)
```




```r
summary(Somerville$costCon)
```

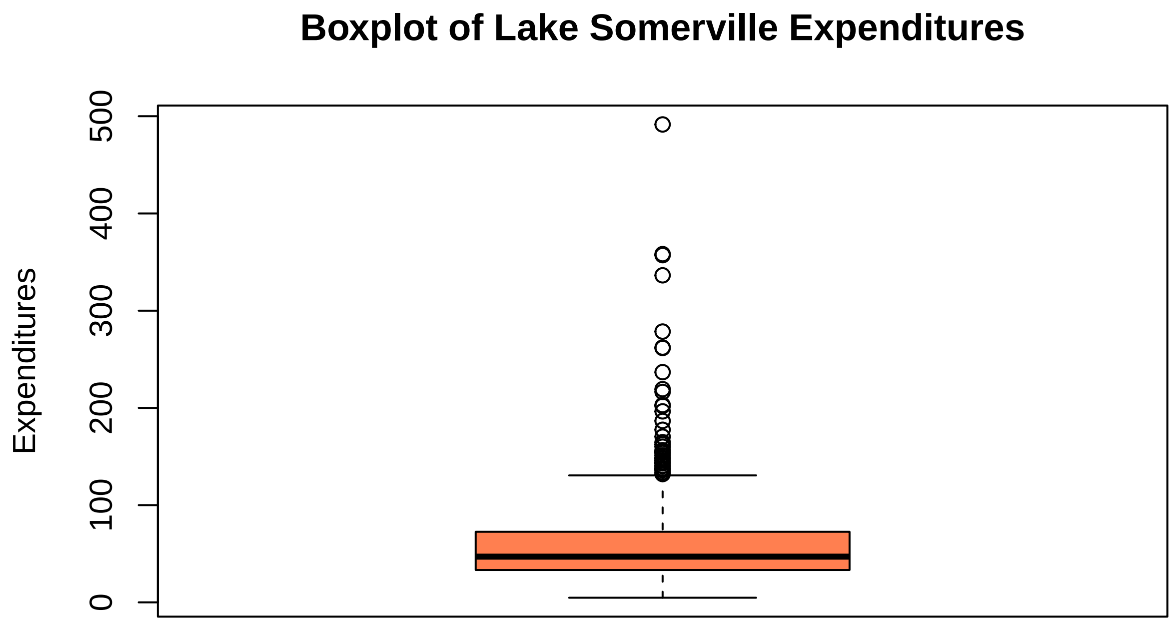
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.34   28.24   41.19   55.42   69.67  493.77
```

This distribution is heavily right-skewed, indicating that most respondents spent lesser when visiting lake Conroe. The bulk of responses are concentrated under 100, with the frequency decreasing sharply as the cost increases. There are also a few outliers on the higher end. It can be observed that some individuals perceive a much higher cost due to the tail.

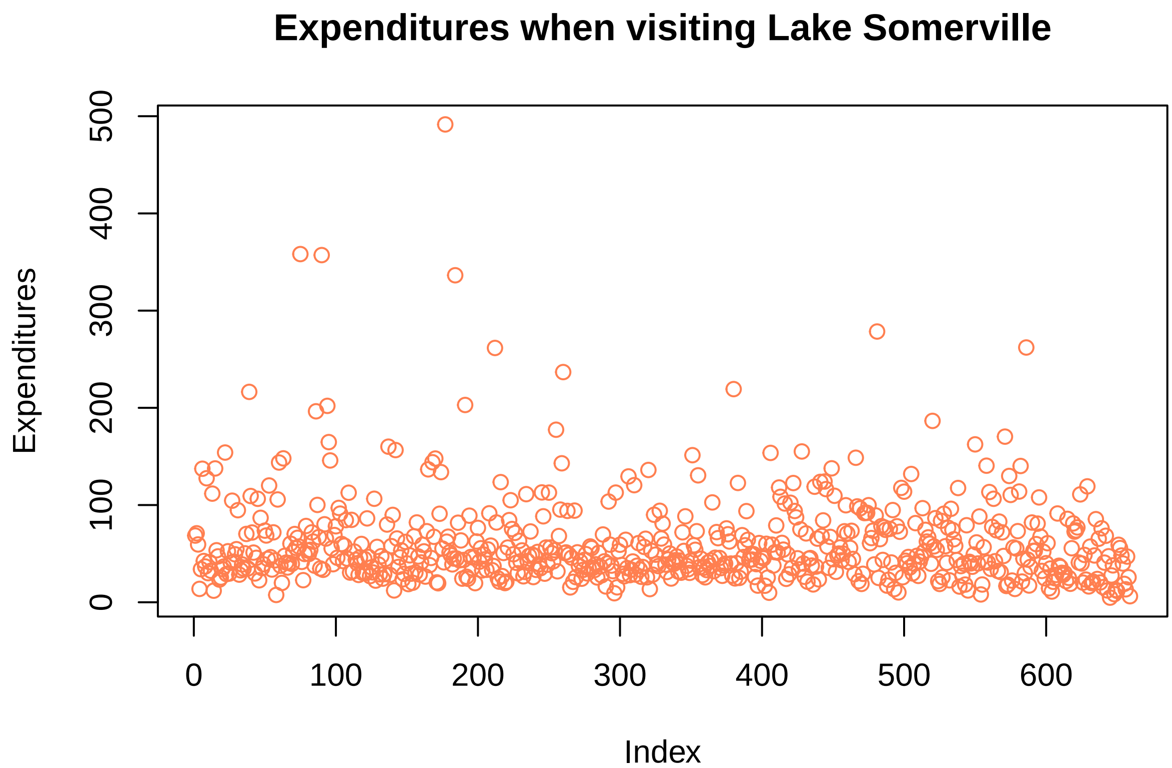
The boxplot shows that most expenditures fall between approximately $20 and $75, with a median around $40. There are many outliers above the upper whisker, some exceeding $400, indicating a positively skewed distribution.

The scatterplot illustrates that most visitors spend under $100, with a high concentration below $50. A few data points are significantly higher, again confirming the presence of outliers and a right-skewed pattern.

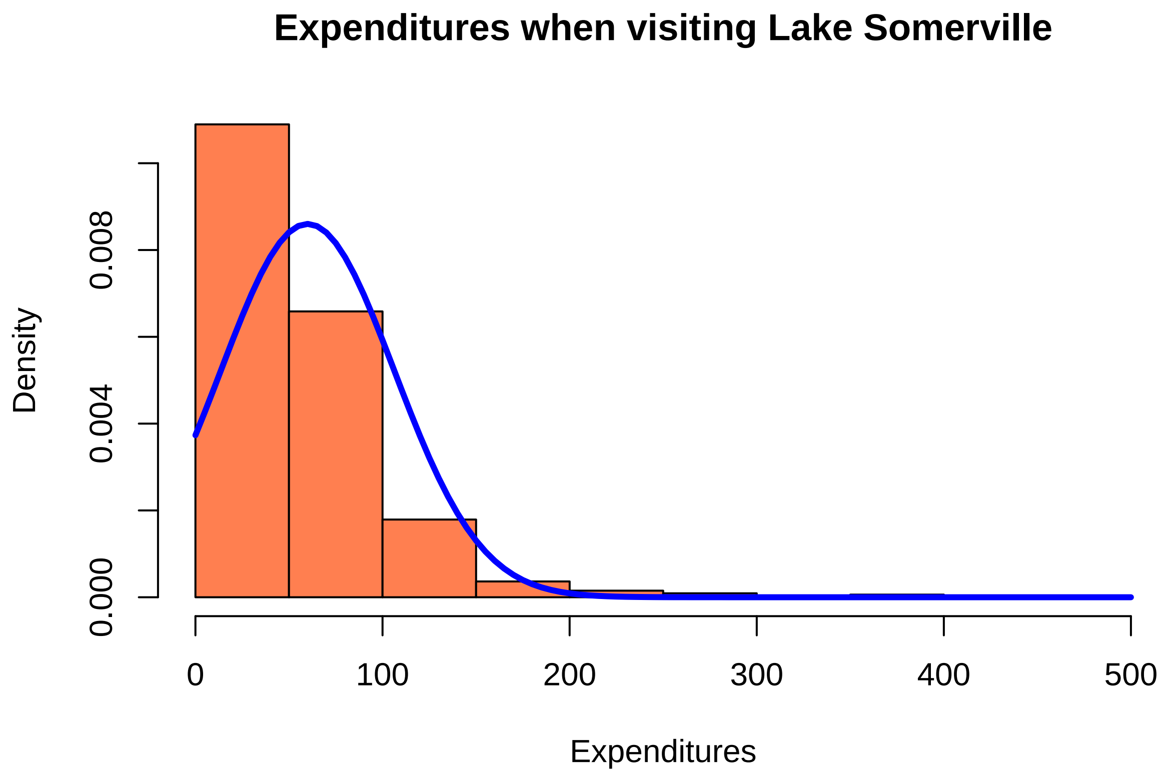
```r
boxplot(Somerville$costSom, main = "Boxplot of Lake Somerville Expenditures",
        ylab = "Expenditures", col = "coral")
```




```r
plot(Somerville$costSom,col = "coral",
     main = "Expenditures when visiting Lake Somerville", ylab = "Expenditures")
```




```r
hist(Somerville$costSom, main = "Expenditures when visiting Lake Somerville",
     xlab = "Expenditures", col = "coral", probability = TRUE)
curve(dnorm(x, mean = mean(Somerville$costSom, na.rm = TRUE),
      sd = sd(Somerville$costSom, na.rm = TRUE)), col = "blue", lwd = 3, add = TRUE)
```




```r
summary(Somerville$costSom)
```

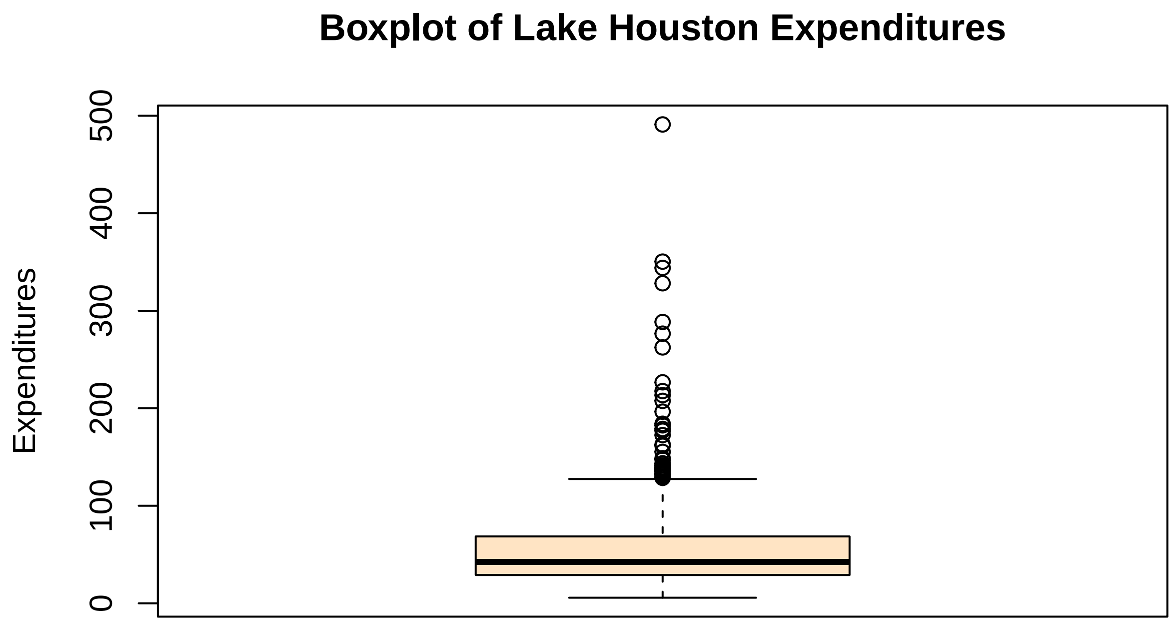
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.767  33.312  47.000  59.928  72.573 491.547
```

This distribution is heavily right-skewed, indicating that most respondents spent lesser when visiting lake Somerville as well. The bulk of responses are concentrated under 100, with the frequency decreasing sharply as the cost increases. There are also a few outliers on the higher end. It can be observed that some individuals perceive a much higher cost due to the tail.

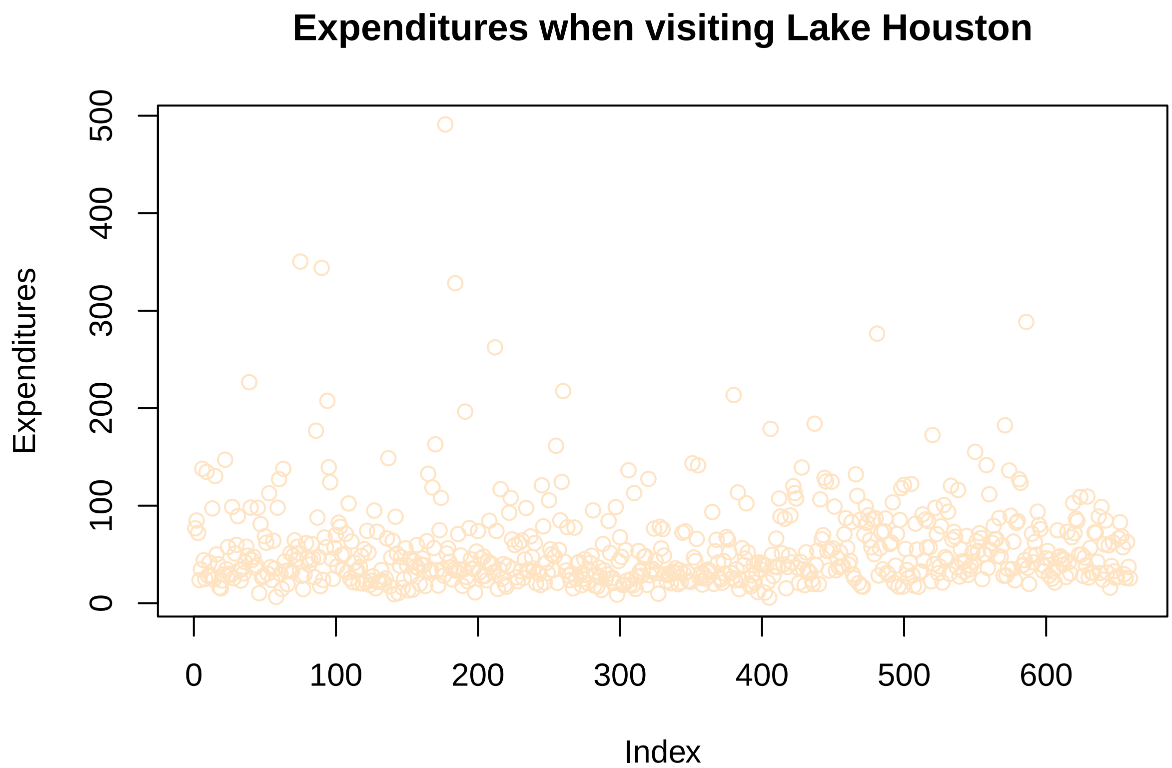
Lake Somerville's boxplot displays a median around $50 and a tight interquartile range. However, it also shows a high number of outliers, with a few expenditures nearing $500.

The scatterplot for Lake Somerville demonstrates a similar pattern of concentrated low expenditures, with a handful of extreme values. Most data points cluster below $100, showing a consistent trend with the boxplot.

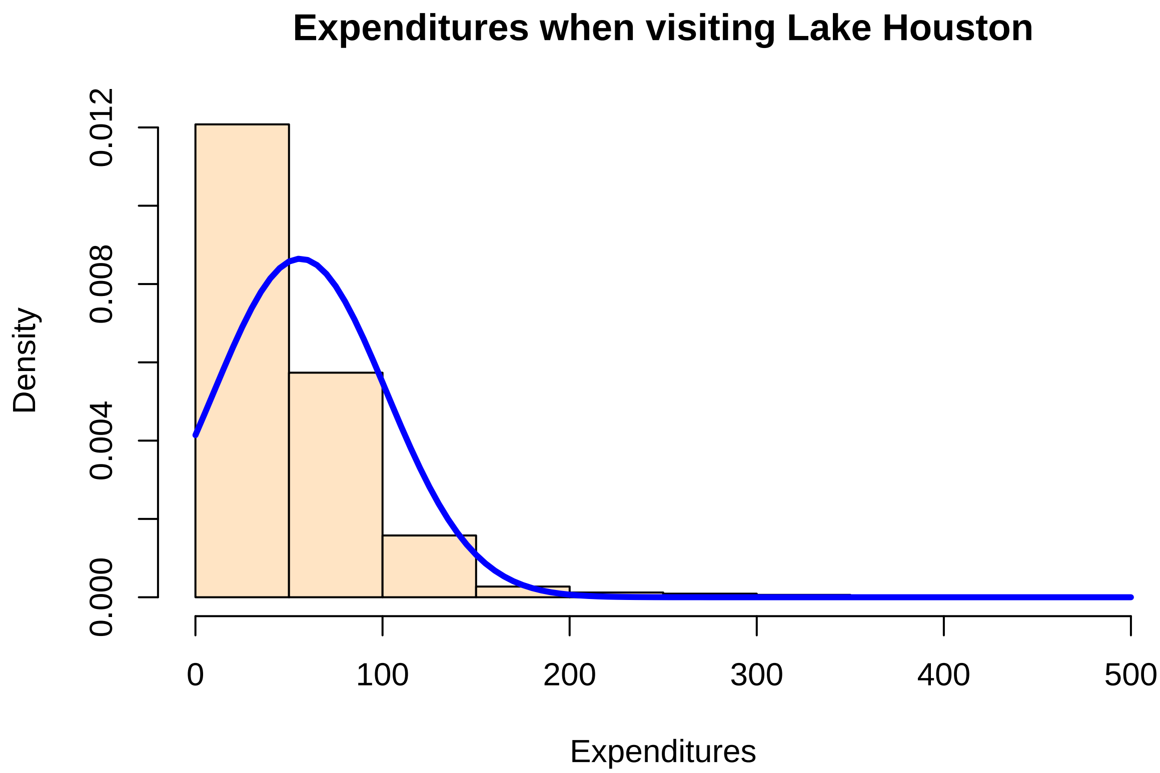
```r
boxplot(Somerville$costHoust, main = "Boxplot of Lake Houston Expenditures",
        ylab = "Expenditures", col = "bisque")
```




```r
plot(Somerville$costHoust,col = "bisque",
     main = "Expenditures when visiting Lake Houston", ylab = "Expenditures")
```




```r
hist(Somerville$costHoust, main = "Expenditures when visiting Lake Houston",
     xlab = "Expenditures", col = "bisque", probability = TRUE)
curve(dnorm(x, mean = mean(Somerville$costHoust, na.rm = TRUE),
      sd = sd(Somerville$costHoust, na.rm = TRUE)), col = "blue", lwd = 3, add = TRUE)
```




```
summary(Somerville$costHoust)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.70   28.96   42.38   55.99   68.56  491.05
```

This distribution is heavily right-skewed, indicating that most respondents spent lesser when visiting lake Houston. The bulk of responses are concentrated under 100, with the frequency decreasing sharply as the cost increases. There are also a few outliers on the higher end. It can be observed that some individuals perceive a much higher cost due to the tail.

The boxplot shows that most expenditures at Lake Houston fall within a narrow range, with the majority clustered below $100. However, there are numerous outliers extending well above this range, indicating occasional high spending by some visitors.

The scatterplot displays individual expenditures, revealing a dense cluster of points at lower spending levels with sporadic high values scattered throughout. This suggests that while most visits incur modest costs, a few visitors spend significantly more.

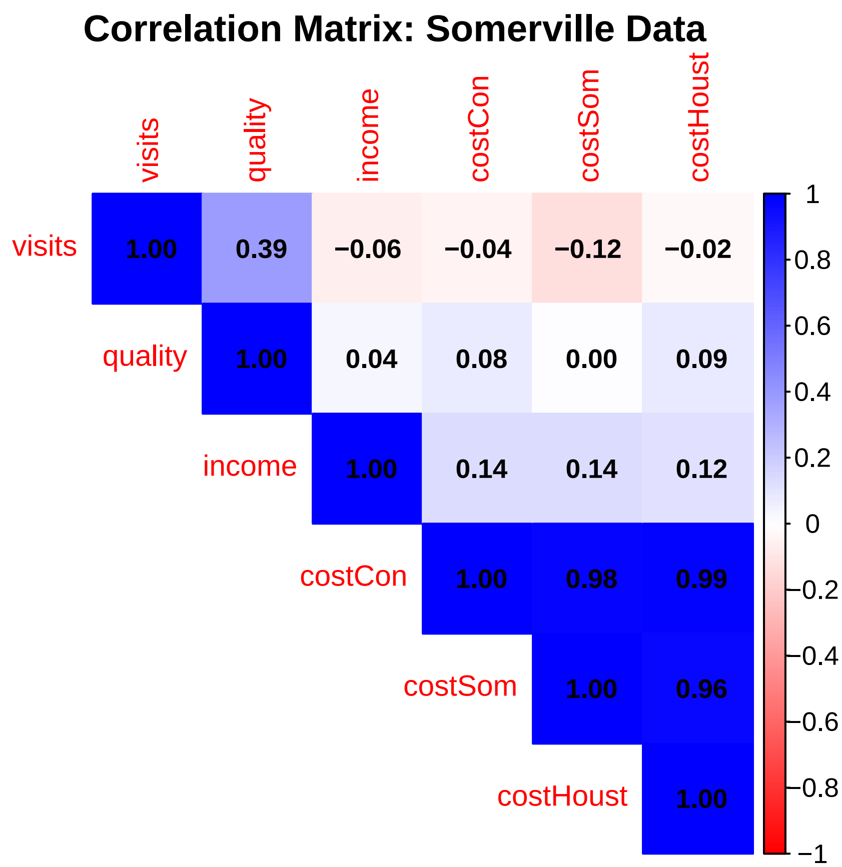
Similar histograms for expenditures on all three lakes indicates that the spending behavior of the respondents across all three lakes isn't that different.

```
data(Somerville, package = "Ecdat")

vars <- Somerville[, c("visits", "quality", "income", "costCon", "costSom", "costHoust")]

cor_matrix <- cor(vars, use = "complete.obs")
```

```r
corrplot(cor_matrix,
         method = "color",        # colorful heatmap
         type = "upper",          # only show upper triangle
         addCoef.col = "black",   # add correlation coefficients
         tl.cex = 0.9,            # text size for variable labels
         number.cex = 0.8,        # text size for coefficients
         col = colorRampPalette(c("red", "white", "blue"))(200),
         title = "Correlation Matrix: Somerville Data",
         mar = c(0,0,1,0))
```



Observations

Visits and quality have a moderate positive correlation (0.39), suggesting that better quality may be associated with more visits.

Visits have very low or negligible correlation with income (-0.06) and the cost variables (all near 0), indicating those factors don't strongly relate to how many visits occur.

Income has a weak positive correlation with cost variables (~0.12 to 0.14), which makes sense—higher income areas might tolerate slightly higher costs.

The cost variables (costCon, costSom, costHoust) are very highly correlated with each other (0.96 to 0.99), indicating they likely move together and could cause multicollinearity issues if used in a regression model together.

(d) Possible Violations

Possible violations of Regression Assumptions:

1. Mutlicollinearity - The variables costCon, costSom, and costHoust are highly correlated with each other, suggesting multicollinearity risks
2. Normality Assumption - All histograms are rightly skewed suggesting violation of the normality assumption
3. Heteroskadsticity - High variablity in dispersion as can be seen from the histograms could affect the error terms
4. Linearity - Weak or non-zero correlations between some variables suggest potential non-linear relationships

(e) Split dataset into training set and test set

```r
set.seed(123)

#Calculate sample size for 80%
sample_size <- floor(0.8 * nrow(Somerville))

#Randomly sample row indices
train_indices <- sample(seq_len(nrow(Somerville)), size = sample_size)

#Create the training (80%) and test (20%) sets
training <- Somerville[train_indices, ]
test <- Somerville[-train_indices, ]
```

Question 3: Multiple Linear Regression Model

```r
# Baseline multiple linear regression model
baseline_model <- lm(visits ~ quality + income + costCon + costSom + costHoust, data = training)

# View summary of the model
summary(baseline_model)
```

```
##
## Call:
## lm(formula = visits ~ quality + income + costCon + costSom +
##     costHoust, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.916  -2.261  -0.442   0.636  80.241
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  2.94399    0.70058   4.202 3.11e-05 ***
## quality      1.03987    0.15181   6.850 2.09e-11 ***
## income      -0.14190    0.14520  -0.977  0.32889
## costCon      0.07137    0.03913   1.824  0.06869 .
## costSom     -0.18705    0.02711  -6.901 1.51e-11 ***
## costHoust    0.10116    0.03226   3.135  0.00181 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.841 on 521 degrees of freedom
## Multiple R-squared:  0.2274, Adjusted R-squared:   0.22
## F-statistic: 30.67 on 5 and 521 DF,  p-value: < 2.2e-16
```

```r
# Make predictions using the baseline model
predictions <- predict(baseline_model, newdata = test)

# Actual values
actual <- test$visits

# Mean Absolute Error (MAE)
mae <- mean(abs(predictions - actual))

# Mean Squared Error (MSE)
mse <- mean((predictions - actual)^2)

print(mse)
```

```
## [1] 15.30251
```

a) Commentary on statistical and economic significance of variables

intercept: The intercept for our data set signifies how many annual visits people would make to Lake Somerville if all of the other predictors equalled zero. This means that, assuming all other predictors equalled zero, individuals would visit Lake Somerville about 2.94 times annually. According to the regression model, the intercept is statistically significant because it has a p-valye of 3.11 x 10^-5, which is far less than the standard alpha level of 0.05. However, there are grounds to believe that the intercept is not economically realistic. For instance, it is unreasonable to think that individuals would visit Lake Somerville if the quality rating of the Lake were 0 or if their incomes were 0. Therefore, while the intercept is statistically significant, it is not economically realistic in this instance.

quality: This variable indicates that, holding all other variables constant, if the quality ranking of Lake Somerville increases by 1 point, the number of annual visits to the Lake will increase by 1.04. According to our regression model, this variable is stastically significant because it has a p-value of 2.09 x 10^-11, which is far less than the standard alpha level of 0.05. We would also expect this variable to be economically significant because it is reasonable to assume that as the quality of a good (or in this case, a location) increases, the more people would be willing to buy that good (or visit that location). Thus, it makes economic sense that the quality ranking score is statistically significant and that it is positively correlated with the number of annual visits to Lake Somerville.

income: This variable indicates that, holding all other variables constant, if the income of an individual increases by 1 unit, the number of annual visits to Lake Somerville that they make will decreases by 0.14. At first glance, this seems counterintuitive because we would expect that people would want to visit Lake Somerville more as their income increases and they have more money to spend. But due to the generally low quality ranking of Lake Somerville, it actually does make economic sense for individuals to visit the Lake less as they make more money. Because they are making more money, they can afford to visit lakes with higher quality ranking scores than Lake Somerville. As individuals' income increases, they visit Lake Somerville less because they are likely choosing to go to other lakes that may be more expensive but are of higher quality. Thus, it makes sense for income to be negatively correlated with the number of annual visits. Although we expect income to be economically significant, our regression model indicates that it is not statistically significance, evidenced by its high p-value of 0.32889. This p-value is higher than the standard alpha level of 0.05, signifyign that income is not statistically significant in the baseline model.

costCon: This variable indicates that a 1 unit increase in the cost of visiting Lake Conroe is associated with 0.071 more visits to Lake Somerville, holding other factors constant. With a p-value of 0.06869, this relationship is not statistically significant. However, economically it would be significant because as the cost of a substitute lake rises, individuals may shift their recreation to Lake Somerville instead. Thus, a positive correlation between costCon and visits to Lake Somerville is reasonable.

costSom: This variable indicates that a 1 unit increase in the cost of visiting Lake Somerville is associated with 0.187 fewer visits to the lake. This result is highly statistically significant ($p < 0.001$) and aligns with economic theory — as the cost of a good increases, demand for it decreases. Thus, it makes sense that costSom and visits to Lake Somerville are negatively correlated.

costHoust: This variable indicates that a 1 unit increase in the cost of visiting Lake Houston, another nearby substitute, is associated with 0.101 more visits to Lake Somerville. This effect is statistically significant ($p < 0.01$), and again, supports economic reasoning — when the price of a substitute good increases, consumers shift their demand toward the relatively cheaper option. Thus, just as with costCon, it makes sense that costHoust and visits to Lake Somerville are positively correlated.

b) Comment on overall fit and significance

The adjusted R-squared of the model is 0.22, indicating that approximately 22% of the variation in the number of visits to Lake Somerville can be explained by the included predictors: quality, income, costCon, costSom, and costHoust. While this is a moderate level of explanatory power and suggests that a significant portion of the variation remains unexplained, this is not uncommon in models involving individual-level behavior. The overall F-statistic is 30.67 with a p-value $< 2.2e-16$, indicating that the model is statistically significant as a whole. This means that at least one of the predictors is significantly associated with the outcome variable.

However, the concerns raised in 2(d) may interfere with the reliability of this fit:

Multicollinearity between cost variables may reduce the precision of individual coefficient estimates, affecting interpretation even if the overall model is significant.

Violations of normality and heteroskedasticity can undermine the assumptions required for valid inference, potentially biasing standard errors and test statistics.

Non-linearity between predictors and the outcome may mean that the linear model is missing important relationships, further limiting explanatory power.

Therefore, while the model is statistically significant, we should explore further regression models that do not violate the assumptions of linear regression for a more reliable and possibly a better fitting model.

Question 4: VIF Test for Multicollinearity

```r
install.packages("car", repos = "https://cloud.r-project.org/")
```

```
##
```

```
## The downloaded binary packages are in
##  /var/folders/sh/y5krnstj3pn0wqd0wr_prc3m0000gn/T//RtmpcqjhCU/downloaded_packages
```

```r
library(car)
```

```
## Loading required package: carData
```

```r
vif(baseline_model)
```

```
##   quality    income   costCon   costSom costHoust
##  1.143402  1.032863 53.088780 24.837765 35.478698
```

From the VIF values for each variable, costCon, costSom and costHoust are much greater than 5, suggesting that multicollinearity is present. This is unsurprising, as the correlation matrix showed similar results. Let's try excluding costCon (highest VIF score) from the model and seeing if that makes a difference.

```r
# Baseline multiple linear regression model
VIF_adjusted_model1 <- lm(visits ~ quality + income + costSom + costHoust, data = training)

# View summary of the model
summary(VIF_adjusted_model1)
```

```
##
## Call:
## lm(formula = visits ~ quality + income + costSom + costHoust,
##     data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.190  -2.249  -0.437   0.698  80.761
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.64177    0.68223   3.872 0.000122 ***
## quality      1.07507    0.15092   7.124 3.51e-12 ***
## income      -0.12048    0.14505  -0.831 0.406582
## costSom     -0.15825    0.02208  -7.166 2.65e-12 ***
## costHoust    0.14417    0.02207   6.532 1.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.854 on 522 degrees of freedom
## Multiple R-squared:  0.2225, Adjusted R-squared:  0.2165
## F-statistic: 37.34 on 4 and 522 DF,  p-value: < 2.2e-16
```

```r
# View VIF scores of new model
vif(VIF_adjusted_model1)
```

```
##   quality    income   costSom costHoust
##  1.124925  1.026105 16.413089 16.528726
```

These VIF values seem to be much lower than the previous ones. This supports our decision of removing costCon from the model in the previous step. Now, as costSom and costHoust still have VIF values greater than 5, let's remove costHoust, which has the higher VIF value out of the two.

```r
# Baseline multiple linear regression model
VIF_adjusted_model2 <- lm(visits ~ quality + income + costSom, data = training)

# View summary of the model
summary(VIF_adjusted_model2)
```

```
##
## Call:
## lm(formula = visits ~ quality + income + costSom, data = training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.003 -1.530 -0.717  0.105 83.526
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.190776   0.705247   3.106  0.00200 **
## quality      1.395723   0.148289   9.412  < 2e-16 ***
## income      -0.196692   0.150227  -1.309  0.19101
## costSom     -0.018523   0.005706  -3.246  0.00124 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.083 on 523 degrees of freedom
## Multiple R-squared:  0.1589, Adjusted R-squared:  0.1541
## F-statistic: 32.94 on 3 and 523 DF,  p-value: < 2.2e-16
```

```r
# View VIF scores of new model
vif(VIF_adjusted_model2)
```

```
##  quality   income  costSom
## 1.005919 1.019465 1.015019
```

Now all variables have VIF values smaller than 5, so we will proceed with this model in the next steps.

Question 5: AIC Test for Model Fit

```r
# Using AIC to determine the subset of predictors
AIC(baseline_model)
```

```
## [1] 3363.705
```

```r
AIC(VIF_adjusted_model2)
```

```
## [1] 3404.471
```

```r
summary(baseline_model)
```

```
##
## Call:
## lm(formula = visits ~ quality + income + costCon + costSom +
##     costHoust, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.916  -2.261  -0.442   0.636  80.241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.94399    0.70058   4.202 3.11e-05 ***
## quality      1.03987    0.15181   6.850 2.09e-11 ***
## income      -0.14190    0.14520  -0.977  0.32889
## costCon      0.07137    0.03913   1.824  0.06869 .
## costSom     -0.18705    0.02711  -6.901 1.51e-11 ***
## costHoust    0.10116    0.03226   3.135  0.00181 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.841 on 521 degrees of freedom
## Multiple R-squared:  0.2274, Adjusted R-squared:   0.22
## F-statistic: 30.67 on 5 and 521 DF,  p-value: < 2.2e-16
```

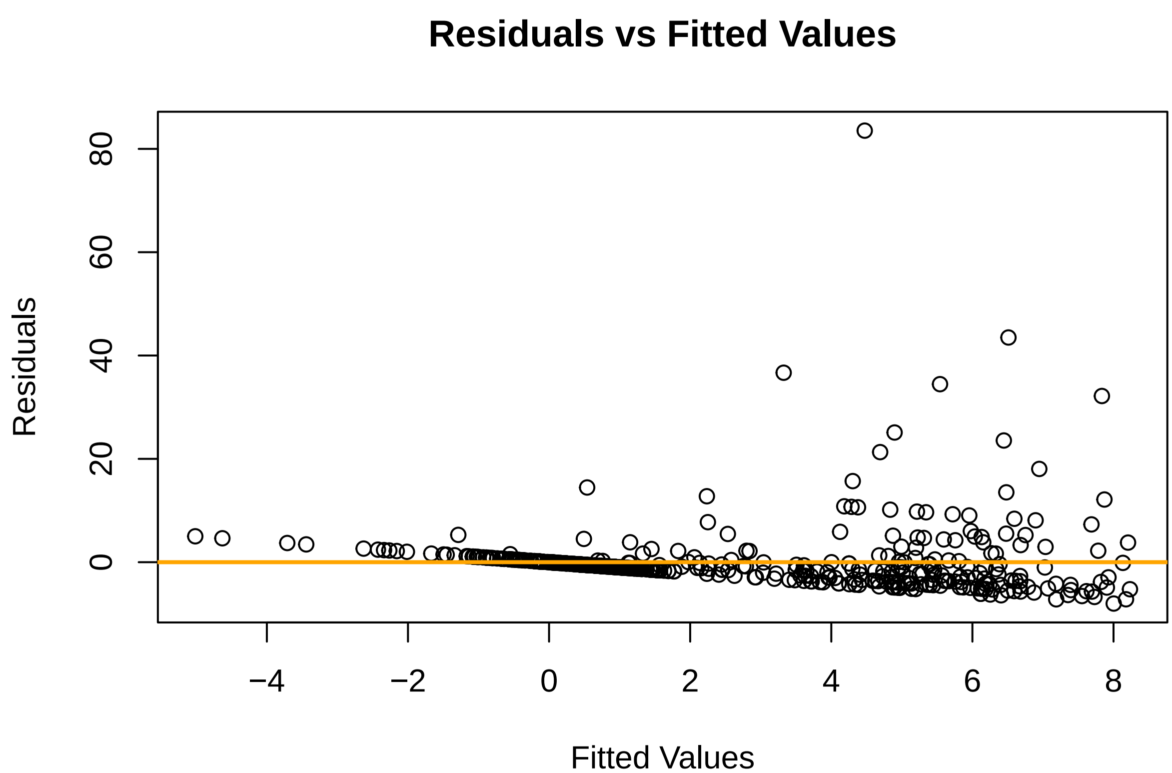
```r
summary(VIF_adjusted_model2)
```

```
##
## Call:
## lm(formula = visits ~ quality + income + costSom, data = training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.003 -1.530 -0.717  0.105 83.526
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.190776   0.705247   3.106  0.00200 **
## quality      1.395723   0.148289   9.412  < 2e-16 ***
## income      -0.196692   0.150227  -1.309  0.19101
## costSom     -0.018523   0.005706  -3.246  0.00124 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.083 on 523 degrees of freedom
## Multiple R-squared:  0.1589, Adjusted R-squared:  0.1541
## F-statistic: 32.94 on 3 and 523 DF,  p-value: < 2.2e-16
```

To find the best set of predictors, we used stepwise selection based on the Akaike Information Criterion (AIC), starting from the VIF-adjusted model from step 4. This gave us a model with an AIC of 3404. Although this is a bit higher than the AIC of our original model in question 3 (3363.705), we chose to start from the VIF-adjusted model to fix multicollinearity — where predictors are too closely related and can make the results less reliable. So even though the AIC went up a little, this model is more trustworthy. Compared to the original model, the VIF adjusted model is simpler, with fewer variables and less overlap. While the fit is not as strong, this tradeoff helps make the model easier to interpret and more reliable.

Question 6: Residuals vs Fitted Values Plot

```r
plot(VIF_adjusted_model2$fitted.values, VIF_adjusted_model2$residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values")
abline(h = 0, col = "orange", lwd = 2)
```



The residuals vs. fitted values plot shows a clear pattern of increasing variance as fitted values increase, indicating heteroskedasticity. This violates an important assumption of linear regression and suggests that the model errors are not evenly distributed, which may affect the inference.

Question 7: RESET Test for Model Misspecification

```r
# install package
install.packages("lmtest", repos = "https://cloud.r-project.org/")
```

```
##
## The downloaded binary packages are in
##  /var/folders/sh/y5krnstj3pn0wqd0wr_prc3m0000gn/T//RtmpcqjhCU/downloaded_packages
```

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
# RESET Test
resettest(VIF_adjusted_model2)
```

```
##
##  RESET test
##
## data:  VIF_adjusted_model2
## RESET = 5.1987, df1 = 2, df2 = 521, p-value = 0.005814
```

```r
resettest(VIF_adjusted_model2, power = 2)
```

```
##
##  RESET test
##
## data:  VIF_adjusted_model2
## RESET = 3.1611, df1 = 1, df2 = 522, p-value = 0.07599
```

```r
resettest(VIF_adjusted_model2, power = 3)
```

```
##
##  RESET test
##
## data:  VIF_adjusted_model2
## RESET = 4.3897, df1 = 1, df2 = 522, p-value = 0.03664
```

We decided to conduct three RESET tests which each test for model misppecification with regard to whether we need to include higher powered terms. The default RESET test tests for both the need to include quadratic terms and cubic terms while the RESET tests with "power = 2" and "power = 3" are more specific tests which tell us whether we need to include quadratic terms or cubic terms, respectively. The default RESET test conducted for VIF_adjusted_model2 produced a RESET statistic of 5.1987 with a small p-value of 0.005814. Since the p-value is far less than 0.05, we must reject the null nypothesis that the model is correctly specified. This suggests that this model is misspecified, meaning it's likely missing non-linear terms or interaction effects. A model with improved specification would likely include powered terms such as squared or cubed variables, and/or interaction variables.

The "power = 2" RESET test gives us a p-value of 0.07599, which is greater than 0.05. This means we fail to reject our null hypothesis, indicating that our model does not need a quadratic term.

The "power = 3" RESET test gives us a p-value of 0.03664, which is smaller than 0.05. This means we reject our null hypothesis, indicating that our model may benedit from a cubic term.

Question 8: Correcting Heteroskedasticity

Using the Breusch-Pagan Test to Test for Heteroskedasticity

```
bptest(VIF_adjusted_model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  VIF_adjusted_model2
## BP = 8.602, df = 3, p-value = 0.03508
```

The null hypothesis of the BP test states that the error variances are constant (homoskedastic); however, as our p-value < 0.05, we must reject the null. Therefore, we can conclude that heteroskedasticity is present in our model and we must correct it so as to not violate the assumptions of linear regression.

Using Robust Standard Errors to Correct the Heteroskedasticity

```
cov1 <- hccm(VIF_adjusted_model2, type="hc3")
coeftest(VIF_adjusted_model2, vcov.=cov1) #produces the robust standard errors
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  2.1907759  0.7539387  2.9058  0.003819 **
## quality      1.3957233  0.1576074  8.8557 < 2.2e-16 ***
## income      -0.1966922  0.1117145 -1.7607  0.078879 .
## costSom     -0.0185235  0.0056143 -3.2993  0.001035 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By applying robust standard errors, the inferences are now valid, even with heteroskedasticity. We can now move on to incorporating interaction terms and higher order terms.

Question 9: Final Model Selection

```
## Final Models 1 & 2: Compare With VIF Adjusted Model to Find Best Model

model_final1 <-
lm(visits ~ quality * income + costSom + I(costSom^3), data = training)

model_final2 <-
lm(visits ~ quality * income + costSom * income + I(costSom^3) + I(income^3), data = training)

summary(model_final2)
```

```
##
## Call:
## lm(formula = visits ~ quality * income + costSom * income + I(costSom^3) +
##     I(income^3), data = training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.626 -1.569 -0.787  0.301 82.683
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.544e+00  1.272e+00   2.000  0.04601 *
## quality         1.842e+00  3.692e-01   4.989  8.3e-07 ***
## income         -9.494e-02  3.758e-01  -0.253  0.80065
## costSom        -4.375e-02  1.370e-02  -3.194  0.00149 **
## I(costSom^3)    9.682e-08  6.327e-08   1.530  0.12656
## I(income^3)    -2.740e-03  4.020e-03  -0.682  0.49581
## quality:income -1.198e-01  8.708e-02  -1.376  0.16933
## income:costSom  4.153e-03  2.973e-03   1.397  0.16303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.067 on 519 degrees of freedom
## Multiple R-squared:  0.1697, Adjusted R-squared:  0.1585
## F-statistic: 15.15 on 7 and 519 DF,  p-value: < 2.2e-16
```

```r
## AIC Test for Best Model Fit

AIC(baseline_model, VIF_adjusted_model2, model_final1, model_final2)
```

```
##                     df      AIC
## baseline_model       7 3363.705
## VIF_adjusted_model2  5 3404.471
## model_final1         7 3403.788
## model_final2         9 3405.686
```

Model Selection Rationale

We created multiple models because we wanted to explore the relationships between predictors (quality, income, costSom) and the response variable (visits). We tested interaction terms and quadratic terms to capture potential non-linearities and interactions effects. The inclusion of terms like quality * income and I(costSom^3) helps model more complex economic relationships, such as how income's effect on visits may vary with quality.

The interaction term quality * income shows how income affects visits differently depending on quality of the site. The quadratic term I(costSom^3) models how the impact of cost on visits changes as cost increases, reflecting realistic cost sensitivity. These terms help make the model more aligned with economic expectations.

Of all the models we tested, model_final1 had the second lowest AIC, higher than only the baseline model. However, according to the VIF test, there is significant multicollinearity between 3 of the variables included in our baseline model: costSom, costCon, and costHoust. Multicollinearity can inflate the variance of the estimated coefficients, overfit the data, and overall lead to a less reliable model. For these reasons, we've decided to choose model_final1, which only includes costSom to prevent inclusion of redundant information, as our optimal model. Model_final2 led to a higher AIC, indicating overfitting. VIF_adjusted_model2 was too simple and also had a higher AIC. Therefore, we believe model_final1 is our best model.

Checking for Heteroskedasticity

```r
bptest(model_final1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_final1
## BP = 10.82, df = 5, p-value = 0.05508
```

Once again, our p-value < 0.05, so we must reject the null hypothesis. Therefore, we can conclude that heteroskedasticity is present in our model and we must correct it so as to not violate the assumptions of linear regression.

Using Robust Standard Errors to Correct the Heteroskedasticity

```r
cov2 <- hccm(model_final1, type="hc3")
coeftest(model_final1, vcov.=cov2) #produces the robust standard errors
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     2.1548e+00  8.2982e-01  2.5967  0.009677 **
## quality         1.7833e+00  3.2862e-01  5.4266 8.814e-08 ***
## income         -5.7556e-02  7.3546e-02 -0.7826  0.434221
## costSom        -2.8413e-02  1.3800e-02 -2.0589  0.040002 *
## I(costSom^3)    1.1471e-07  2.3959e-07  0.4788  0.632293
## quality:income -9.9191e-02  6.6096e-02 -1.5007  0.134036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 10

```r
# Make predictions using the final model we chose
predictions <- predict(model_final1, newdata = test)
summary(model_final1)
```

```
##
## Call:
## lm(formula = visits ~ quality * income + costSom + I(costSom^3),
##     data = training)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.530 -1.512 -0.792  0.168 82.967
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.155e+00  8.603e-01   2.505 0.012558 *
## quality         1.783e+00  3.659e-01   4.874 1.45e-06 ***
## income         -5.756e-02  1.853e-01  -0.311 0.756212
## costSom        -2.841e-02  8.025e-03  -3.541 0.000435 ***
## I(costSom^3)    1.147e-07  6.203e-08   1.849 0.064997 .
## quality:income -9.919e-02  8.580e-02  -1.156 0.248174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.067 on 521 degrees of freedom
## Multiple R-squared:  0.1664, Adjusted R-squared:  0.1584
## F-statistic: 20.79 on 5 and 521 DF,  p-value: < 2.2e-16
```

```r
# Actual values
actual <- test$visits

# Mean Absolute Error (MAE)
mae <- mean(abs(predictions - actual))

# Mean Squared Error (MSE)
mse <- mean((predictions - actual)^2)

# Print results
cat("Mean Absolute Error (MAE):", round(mae, 2), "\n")
```

```
## Mean Absolute Error (MAE): 2.4
```

```r
cat("Mean Squared Error (MSE):", round(mse, 2), "\n")
```

```
## Mean Squared Error (MSE): 17.41
```

We tested how well our final model (model_final1) performs on the test data, which made up 20% of the full dataset. The Mean Absolute Error (MAE) came out to 2.4, meaning that, on average, our model's predictions were off by about 2.4 units. The Mean Squared Error (MSE) was 17.41, which gives more weight to larger errors.

Overall, this shows that our model does a decent job predicting the outcome. The relatively low MAE tells us that most of the predictions are close to the actual values. However, the higher MSE suggests that there were a few predictions that deviated from the actual values by a lot. The large gap between the MAE and MSE indicates that while the model performs well on average, those few large errors are skewing the overall performance, possibly due to outliers or model misspecification. Even with that, the model holds up well when applied to the test data and gives helpful predictions for our response variable.

Question 11

In conclusion, while the baseline model had a greater number of statistically significant variables and lower MSE, this is likely due to its simpler structure, which may have overlooked important interactions or suffered from omitted variable bias. In contrast, the final model provides a more accurate and robust representation of the data, even if fewer individual predictors remain significant. This highlights the trade-off between simplicity and model reliability—emphasizing that statistical significance alone isn't always the best indicator of model quality. The dataset itself, with its large sample size, did not present limitations for model building.

Ultimately, the number of visits to Lake Somerville is influenced by lake quality, the interaction between quality and income, and the cost of visiting the lake, including non-linear effects of cost.