Final Project

04/10/2024

```r
library(readxl)
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.3.3
```

```r
library(knitr)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3

## corrplot 0.92 loaded
```

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.3.3
```

```r
data <- read_excel("C:/Users/Suma2/Downloads/Applied stats_projrct/cawlifornis.xlsx")


#data <- read_excel("C:/Users/Owner/Desktop/Applied Statistics/Project/TaxRevenue-NJ.xlsx")
```

```r
head(data)
```

```
## # A tibble: 6 x 16
##   State  Year Quarter StateRevenue AvgTaxRate TaxRateRank AvgTaxRateOnWages
##   <chr> <dbl>   <dbl>        <dbl>      <dbl>       <dbl>             <dbl>
## 1 CA     1997       1      3553666       0.88          18              3.55
## 2 CA     1997       2      3270262       0.75          23              3.48
## 3 CA     1997       3      3285079       0.71          26              3.43
## 4 CA     1997       4      3259747       0.68          26              3.39
## 5 CA     1998       1      3269140       0.63          26              3.12
## 6 CA     1998       2      3151684       0.69          25              3.08
## # i 9 more variables: AvgTaxRateOnWagesRank <dbl>, MinTaxWage <dbl>,
## #   TrustFund <dbl>, TFPerWages <dbl>, TFWagesRank <dbl>, Interest <dbl>,
## #   HighCostMultiple <dbl>, AvgHCM <dbl>, AvgHCMRank <dbl>
```

Description of Variables:

- **State:** Represents the state where the data is recorded.
- **Year:** Indicates the calendar year for the data.
- **Quarter:** Represents the quarter (1, 2, 3, or 4) of the year in which the data is recorded.
- **StateRevenue:** Reflects the state revenue over the past 12 months.
- **AvgTaxRate:** Represents the average tax rate over the past 12 months, expressed as a percentage.
- **TaxRateRank:** Indicates the rank of the average tax rate over the past 12 months among other states.
- **AvgTaxRateOnWages:** Reflects the average tax rate on taxable wages over the past 12 months, expressed as a percentage.
- **AvgTaxRateOnWagesRank:** Indicates the rank of the average tax rate on taxable wages over the past 12 months among other states.
- **MinTaxWage:** Represents the taxable wage base, which is the maximum amount of earnings subject to a particular tax.
- **TrustFund:** Reflects the balance in the trust fund.
- **TFPerWages:** Indicates the trust fund balance as a percentage of total wages.
- **TFWagesRank:** Indicates the rank of the trust fund balance among other states based on total wages.
- **Interest:** Represents the interest earned on the trust fund.
- **HighCostMultiple:** Reflects the high cost multiple.
- **AvgHCM:** Represents the average high cost multiple ACHM.
- **AvgHCMRank:** Indicates the rank of the average high cost multiple ACHM among other states.

```
any(is.na(data))
```

```
## [1] TRUE
```

```
final_data <- na.omit(data)
```

```
final_data <- final_data[, -1]
```

Exploratory Data Analysis:

```
names(final_data)
```

```
##  [1] "Year"                  "Quarter"               "StateRevenue"
##  [4] "AvgTaxRate"            "TaxRateRank"           "AvgTaxRateOnWages"
##  [7] "AvgTaxRateOnWagesRank" "MinTaxWage"            "TrustFund"
## [10] "TFPerWages"            "TFWagesRank"           "Interest"
## [13] "HighCostMultiple"      "AvgHCM"                "AvgHCMRank"
```

```r
kable(summary(final_data))
```

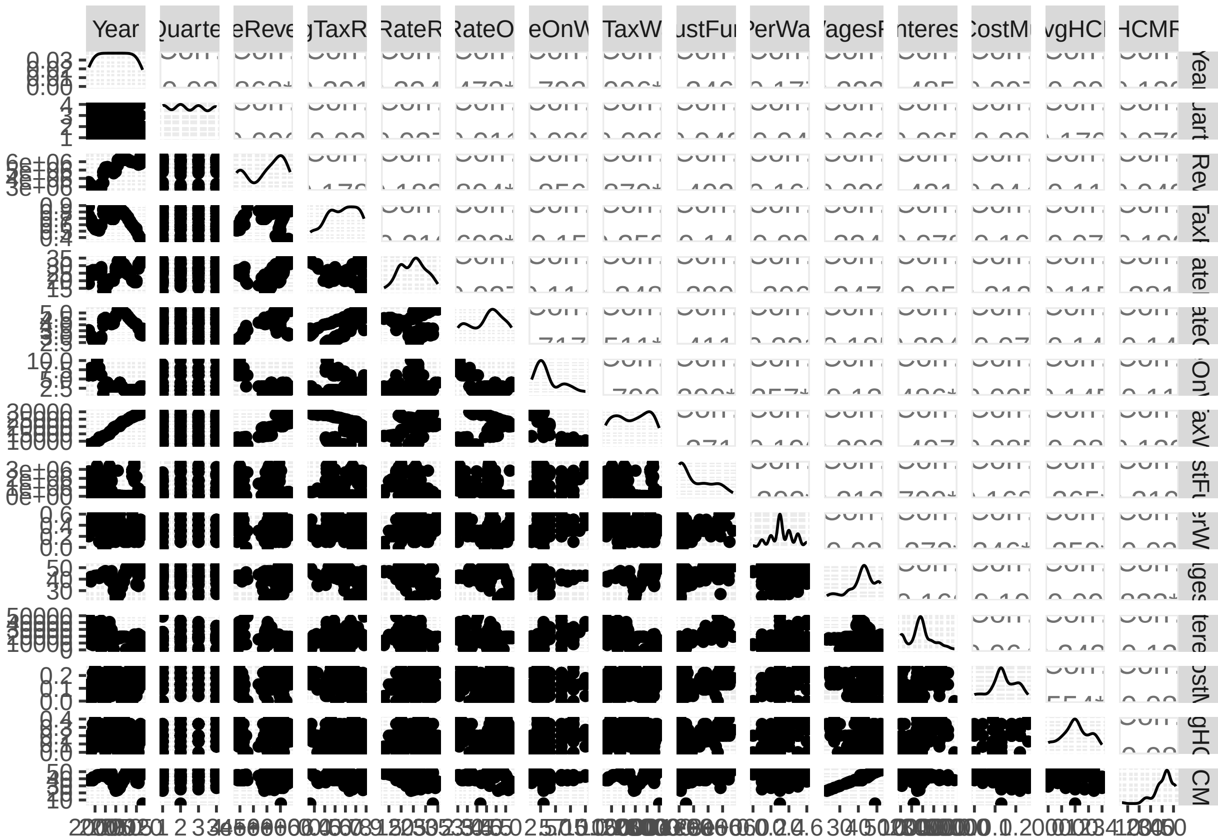
	Year	Quarter	StateRevenueAvgTaxRateTaxRateAvgTaxRateRankAvgTaxRateOnWagesTaxWagesRankTEPerWEWagesIntensityHighCostWageAMHCAvgHCMRank												
	Min. :1997	Min. :1.000	Min. :2741634	Min. :0.3600	Min. :13.00	Min. :2.500	Min. : 1.000	Min. : 7423	Min. : 5949	Min. :0.0000	Min. :23.00	Min. : 0	Min. :0.0000	Min. :0.0000	Min. : 5.00
	1st Qu.:2003	1st Qu.:1.250	1st Qu.:3372818	1st Qu.:0.5600	1st Qu.:20.00	1st Qu.:3.243	1st Qu.: 2.000	1st Qu.:12713	1st Qu.: 77055	1st Qu.:0.3000	1st Qu.:40.00	1st Qu.:12420	1st Qu.:0.1000	1st Qu.:0.1300	1st Qu.:38.00
	Median :2010	Median :2.000	Median :5353483	Median :0.6900	Median :25.00	Median :4.150	Median : 3.000	Median :19524	Median : 622889	Median :0.3000	Median :43.00	Median :18392	Median :0.1300	Median :0.2000	Median :43.00
	Mean :2010	Mean :2.481	Mean :5009545	Mean :0.6675	Mean :24.81	Mean :4.008	Mean : 3.764	Mean :18686	Mean :1107472	Mean :0.3302	Mean :42.34	Mean :17917	Mean :0.1404	Mean :0.2008	Mean :41.88
	3rd Qu.:2016	3rd Qu.:3.000	3rd Qu.:6097644	3rd Qu.:0.7875	3rd Qu.:28.00	3rd Qu.:4.622	3rd Qu.: 6.000	3rd Qu.:25137	3rd Qu.:1900501	3rd Qu.:0.4000	3rd Qu.:46.75	3rd Qu.:22378	3rd Qu.:0.1800	3rd Qu.:0.2700	3rd Qu.:46.75
	Max. :2023	Max. :4.000	Max. :6729503	Max. :0.8800	Max. :35.00	Max. :5.310	Max. :10.000	Max. :30065	Max. :3717242	Max. :0.6000	Max. :52.00	Max. :48727	Max. :0.2600	Max. :0.4000	Max. :52.00

```r
dim(final_data)
```

```
## [1] 106  15
```

- The dimension of the dataset is 85*15

```r
ggpairs(final_data)
```



```r
idx<-which(final_data$TaxRateRank > 20)
idx
```

```
##  [1]   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
## [20]  21  22  23  24  25  26  27  28  44  50  51  52  53  54  55  56  57  58  59
## [39]  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78
## [58]  79  80  81  82  83  84  85  86  97  98  99 100 101 102 103 104 105 106
```

```r
final_data$Year[idx]
```

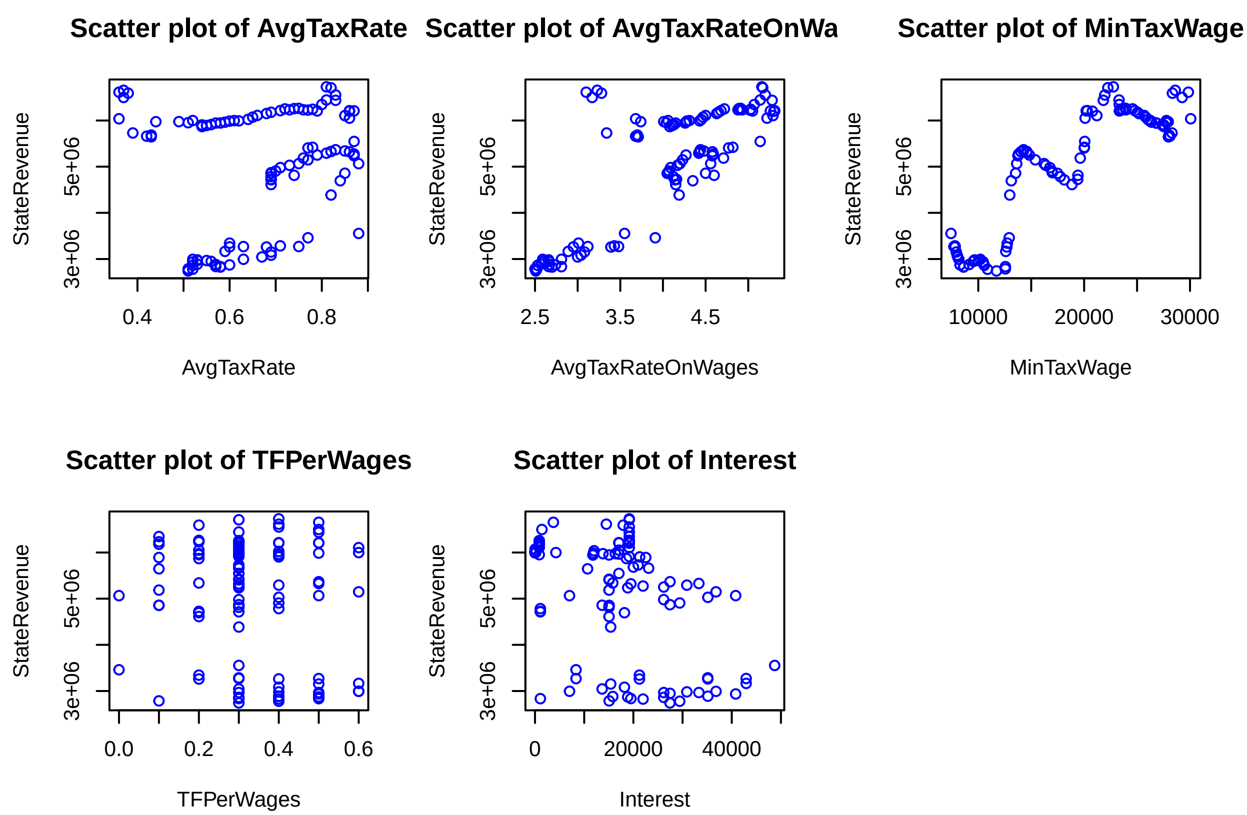
```
##  [1] 1997 1997 1997 1998 1998 1998 1998 1999 1999 1999 1999 2000 2000 2000 2000
## [16] 2001 2001 2001 2001 2002 2002 2002 2002 2003 2003 2003 2003 2007 2009 2009
## [31] 2009 2010 2010 2010 2010 2011 2011 2011 2011 2012 2012 2012 2012 2013 2013
## [46] 2013 2013 2014 2014 2014 2014 2015 2015 2015 2015 2016 2016 2016 2016 2017
## [61] 2017 2017 2017 2018 2018 2021 2021 2021 2021 2022 2022 2022 2022 2023 2023
```

```r
skewness_df <- data.frame(Skewness = sapply(final_data, skewness))
skewness_df
```

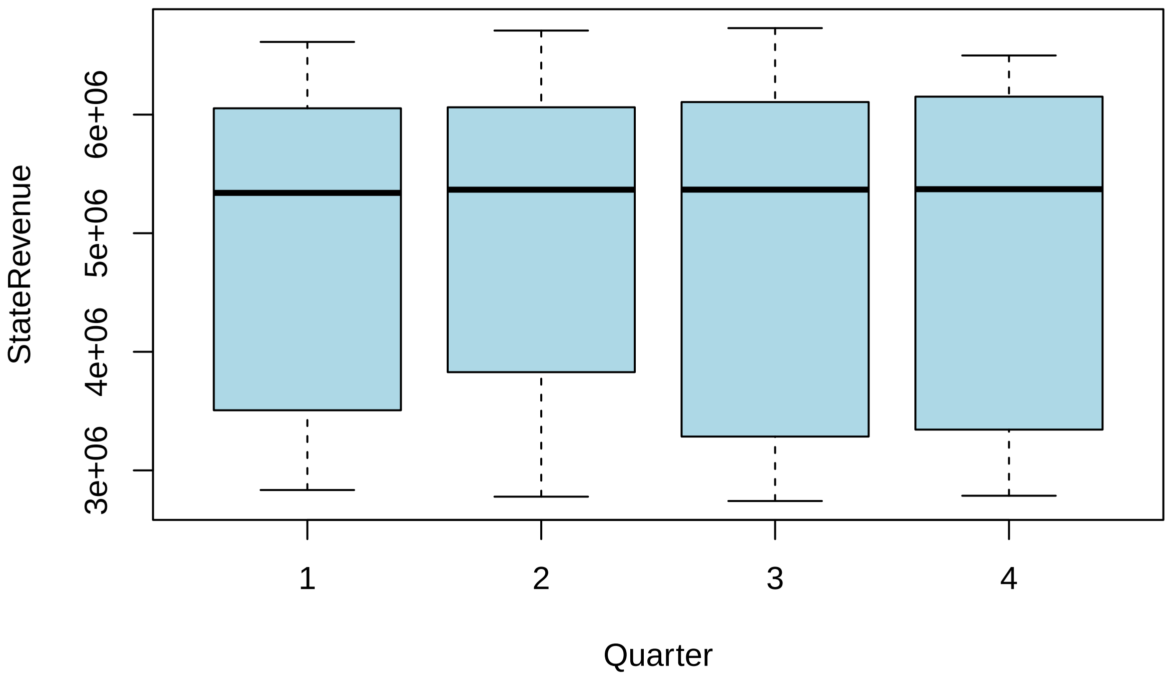
```
##                         Skewness
## Year                 0.003640932
## Quarter              0.026622276
## StateRevenue        -0.590206076
## AvgTaxRate          -0.340367096
```

```
## TaxRateRank               -0.069714033
## AvgTaxRateOnWages         -0.316222967
## AvgTaxRateOnWagesRank      1.054706166
## MinTaxWage                -0.080216599
## TrustFund                  0.644150464
## TFPerWages                -0.039600597
## TFWagesRank               -0.825572446
## Interest                   0.282531430
## HighCostMultiple          -0.290295699
## AvgHCM                    -0.145285270
## AvgHCMRank                -1.304586630
```

```r
feature_names <- c("AvgTaxRate", "AvgTaxRateOnWages", "MinTaxWage", "TFPerWages", "Interest")

par(mfrow = c(2, 3))

for (feature in feature_names) {
  # Create scatter plot with frequencies on y-axis
  plot(final_data[["StateRevenue"]]~final_data[[feature]],
       main = paste("Scatter plot of", feature),
       xlab = feature, ylab = "StateRevenue", col = "blue")
}
#boxplot(StateRevenue~Quarter,data = final_data)
par(mfrow = c(1, 1))
```

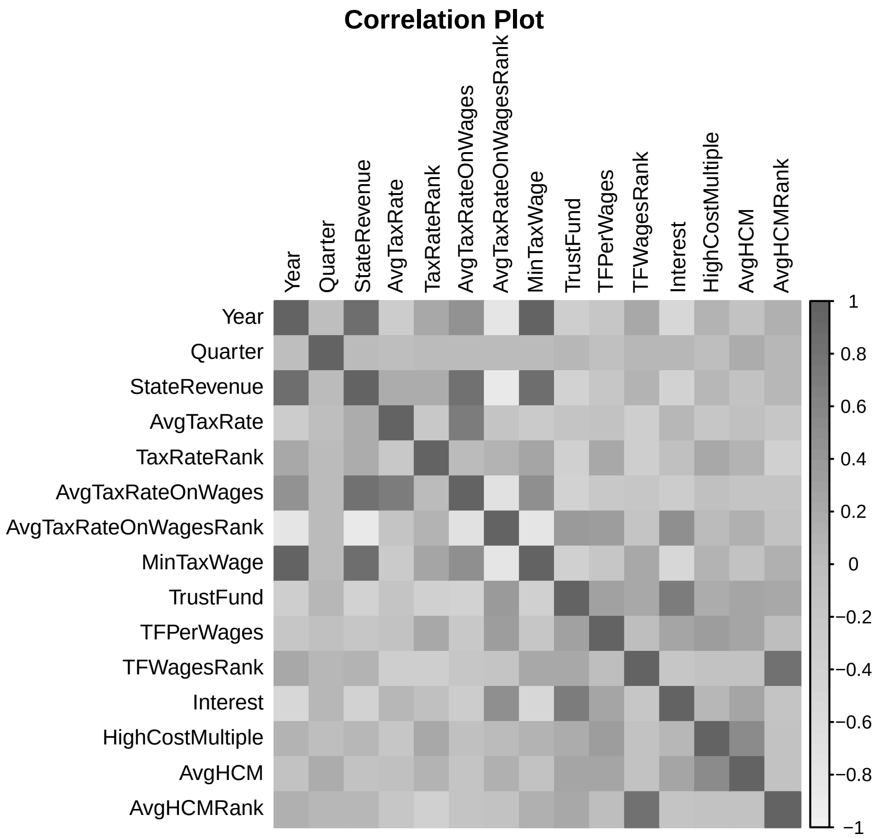


```r
boxplot(StateRevenue~Quarter,data = final_data, col = "lightblue", border = "black")
```

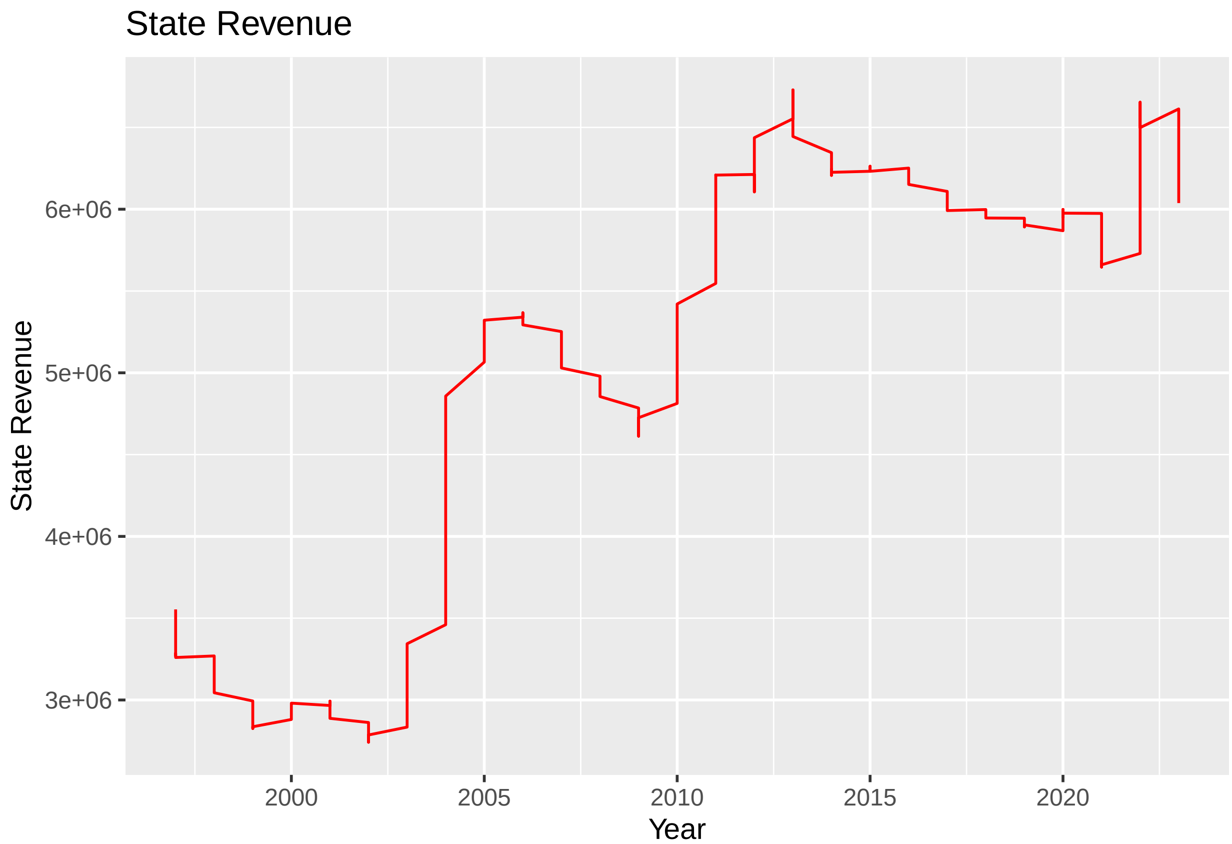



```r
# Reset the layout to default
```

```r
par(mfrow = c(1, 1), cex = 0.7)

cor_matrix <- cor(final_data)
my_colors <- colorRampPalette(c("#f0f0f0", "#bdbdbd", "#636363"))(50)
corrplot(cor_matrix, method = "color", col = my_colors, tl.col = "black", title = "Correlation Plot", ma
```




```r
ggplot(final_data, aes(x = Year)) +
  geom_line(aes(y = StateRevenue), color = "red", linetype = "solid") +
  ggtitle("State Revenue") +
  ylab("State Revenue") +
  xlab("Year")
```




```
set.seed(123)

# Define the proportion of the data you want in the training set
train_proportion <- 0.7

# Generate random indices for the training set
train_indices <- sample(1:nrow(final_data), round(train_proportion * nrow(final_data)))

# Create the training set
train_data <- final_data[train_indices, ]

# Create the testing set excluding the training set
test_data <- final_data[-train_indices, ]
```

REGRESSION:

```
LinearRegresssion_model1 <- lm(StateRevenue ~ AvgTaxRate + TaxRateRank + AvgTaxRateOnWages + AvgTaxRate

summary(LinearRegresssion_model1)
```

```
##
## Call:
## lm(formula = StateRevenue ~ AvgTaxRate + TaxRateRank + AvgTaxRateOnWages +
```

```
##     AvgTaxRateOnWagesRank + MinTaxWage + TrustFund + TFPerWages +
##     TFWagesRank + Interest + HighCostMultiple + AvgHCM + AvgHCMRank,
##     data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -691976 -153509  -44083  101868  830820
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.012e+06  8.158e+05  -1.240    0.220
## AvgTaxRate             9.258e+05  2.540e+06   0.365    0.717
## TaxRateRank            1.767e+04  1.652e+04   1.069    0.289
## AvgTaxRateOnWages      5.268e+05  4.751e+05   1.109    0.272
## AvgTaxRateOnWagesRank -8.319e+04  5.352e+04  -1.554    0.125
## MinTaxWage             1.152e+02  4.651e+01   2.477    0.016 *
## TrustFund             -4.299e-02  9.040e-02  -0.476    0.636
## TFPerWages             5.525e+05  3.781e+05   1.461    0.149
## TFWagesRank            1.968e+04  1.023e+04   1.924    0.059 .
## Interest               1.057e+01  7.771e+00   1.360    0.179
## HighCostMultiple       7.443e+05  7.596e+05   0.980    0.331
## AvgHCM                -6.160e+05  4.778e+05  -1.289    0.202
## AvgHCMRank            -3.043e+03  8.162e+03  -0.373    0.711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 320200 on 61 degrees of freedom
## Multiple R-squared:  0.9524, Adjusted R-squared:  0.9431
## F-statistic: 101.7 on 12 and 61 DF,  p-value: < 2.2e-16
```

The model suggests that certain tax-related factors significantly influence state revenue. Notably, the average tax rate and trust fund size show strong positive relationships with state revenue, indicating that as these increase, so does the revenue. On the other hand, some rankings related to economic distress (like AvgHCMRank) negatively impact revenue. The model is robust, explaining a large proportion of the variability in state revenue and showing strong overall statistical significance.

```r
LinearRegresssion_model2<- lm(StateRevenue ~ Year + Quarter+ AvgTaxRate + TaxRateRank + AvgTaxRateOnWag

summary(LinearRegresssion_model2)
```

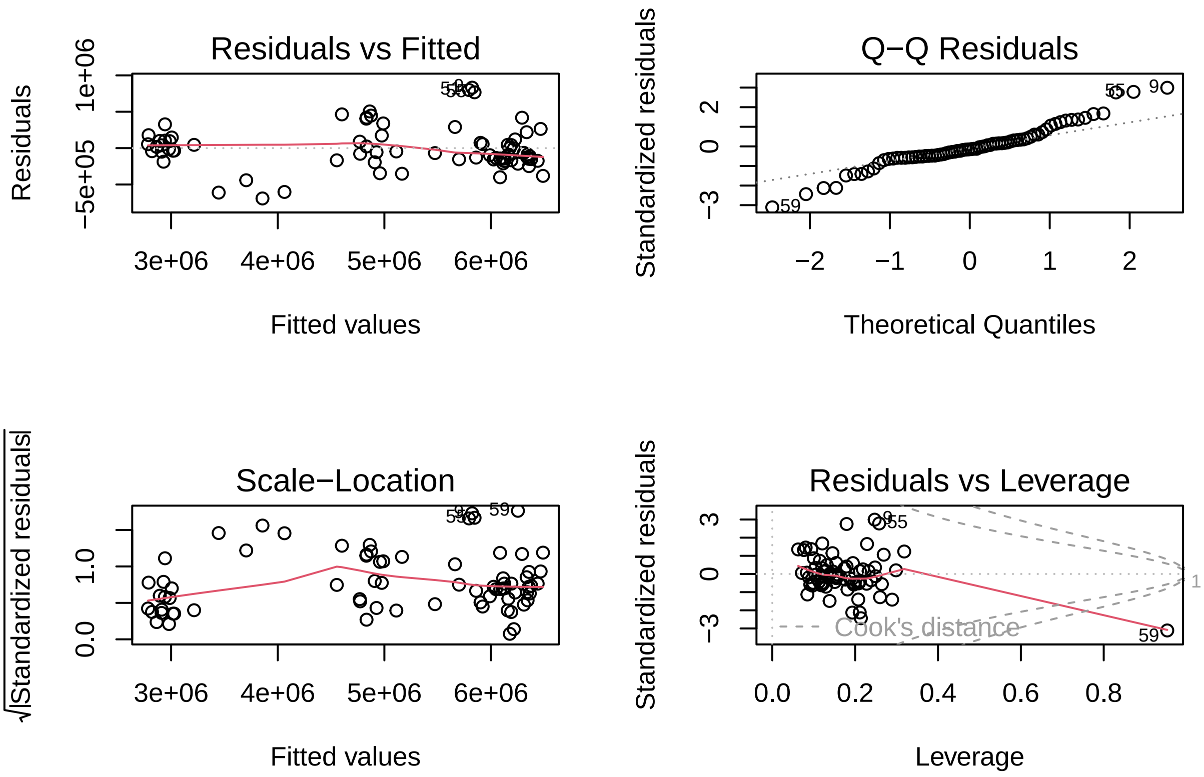
```
##
## Call:
## lm(formula = StateRevenue ~ Year + Quarter + AvgTaxRate + TaxRateRank +
##     AvgTaxRateOnWages + AvgTaxRateOnWagesRank + MinTaxWage +
##     TrustFund + TFPerWages + TFWagesRank + Interest + HighCostMultiple +
##     AvgHCM + AvgHCMRank, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -687679 -125395   -3503  110304  668068
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)            -7.948e+08  1.345e+08  -5.907 1.84e-07 ***
## Year                    3.981e+05  6.747e+04   5.900 1.89e-07 ***
## Quarter                 1.036e+05  3.729e+04   2.778 0.007319 **
## AvgTaxRate              1.448e+06  2.214e+06   0.654 0.515682
## TaxRateRank             1.984e+04  1.346e+04   1.473 0.145944
## AvgTaxRateOnWages       7.013e+05  4.145e+05   1.692 0.095969 .
## AvgTaxRateOnWagesRank  1.755e+04  4.797e+04   0.366 0.715824
## MinTaxWage             -3.019e+02  8.159e+01  -3.700 0.000476 ***
## TrustFund              -3.616e-02  7.323e-02  -0.494 0.623309
## TFPerWages              3.377e+05  3.071e+05   1.100 0.275954
## TFWagesRank             1.048e+04  8.400e+03   1.248 0.216942
## Interest                2.038e+00  6.431e+00   0.317 0.752392
## HighCostMultiple        4.260e+05  6.194e+05   0.688 0.494305
## AvgHCM                 -1.085e+05  4.003e+05  -0.271 0.787408
## AvgHCMRank              4.161e+03  6.802e+03   0.612 0.543117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 258200 on 59 degrees of freedom
## Multiple R-squared:  0.9701, Adjusted R-squared:  0.963
## F-statistic: 136.6 on 14 and 59 DF,  p-value: < 2.2e-16
```

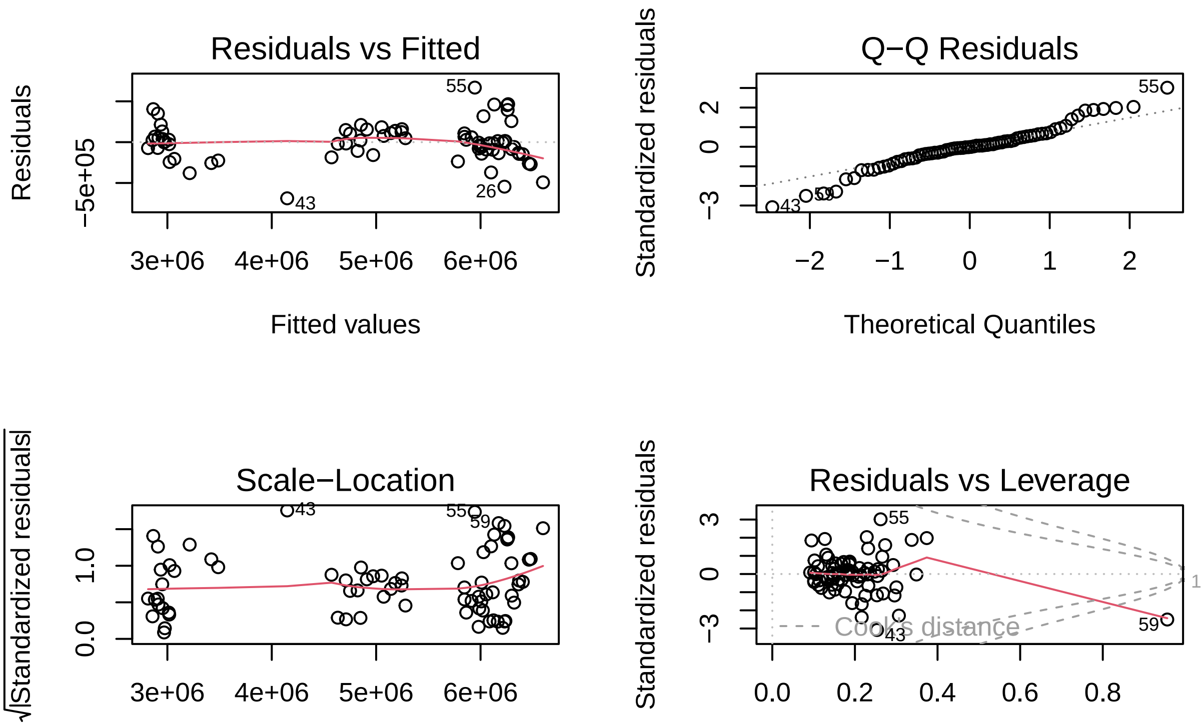
This model is highly effective in explaining state revenue based on the included predictors. Key findings are:

Time Dynamics: Both the Year and Quarter were included to capture time trends and periodic effects, respectively. The year has a significant impact, reflecting perhaps inflationary trends, economic growth, or changes in taxation policy over time. Tax Rates and Economic Indicators: Consistent with economic intuition, higher average tax rates significantly increase state revenue. Economic indicators like the trust fund size also positively affect revenue, while indicators of economic distress (like AvgHCMRank) negatively impact it. The model is statistically strong, and the inclusion of time variables helps account for changes over years and within years, though the latter (quarterly changes) didn't prove significant. This robust model offers a comprehensive view of the factors driving state revenue, making it valuable for forecasting and policy analysis. *Summary of Model1:

```r
par(mfrow = c(2, 2))
plot(LinearRegresssion_model1)
```




```r
par(mfrow = c(2, 2))
plot(LinearRegresssion_model2)
```

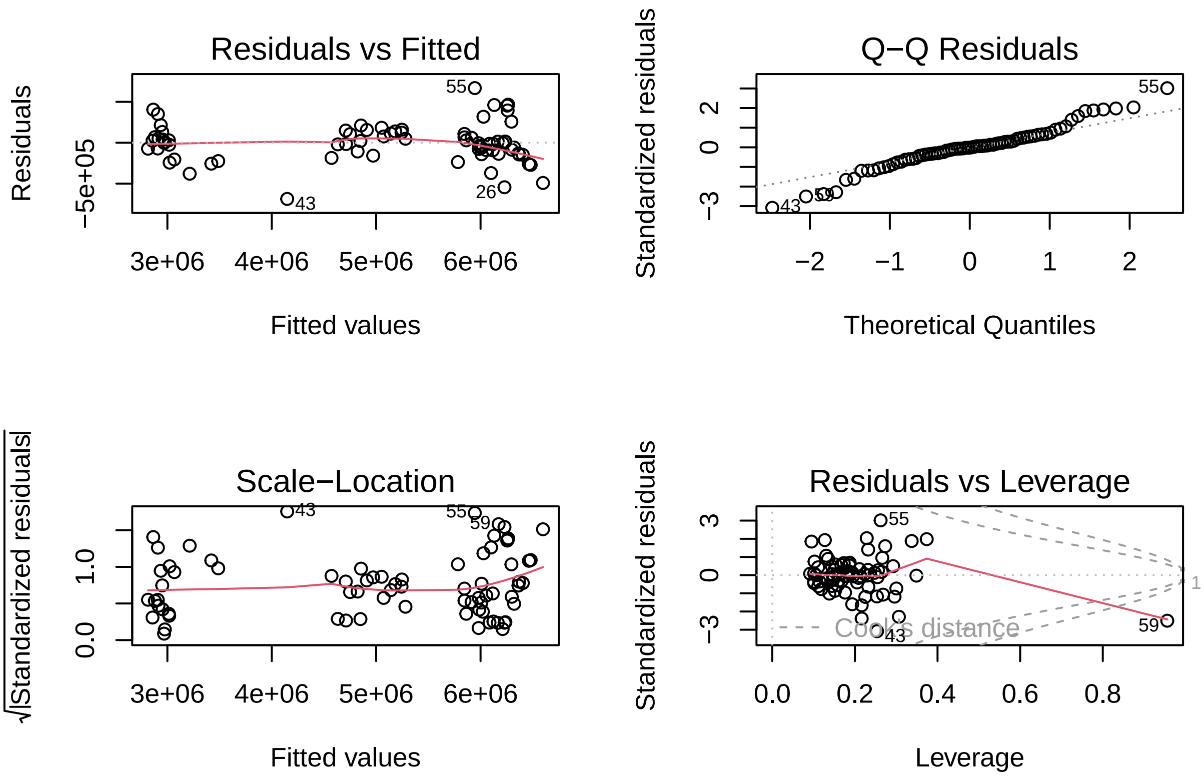



```r
#Backward Selection
modback<-step(lm(final_data$StateRevenue ~ ., data = final_data),  direction = "backward", trace = 0)
summary(modback)
```

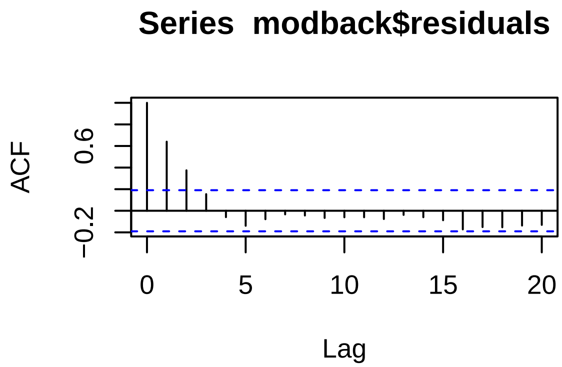
```
##
## Call:
## lm(formula = final_data$StateRevenue ~ Year + Quarter + AvgTaxRate +
##     TaxRateRank + MinTaxWage + TFPerWages + TFWagesRank, data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -816907 -102335  -18610  109992  591803
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.719e+08  8.154e+07  -8.239 7.81e-13 ***
## Year         3.361e+05  4.093e+04   8.212 8.93e-13 ***
## Quarter      9.287e+04  2.278e+04   4.077 9.30e-05 ***
## AvgTaxRate   5.049e+06  2.056e+05  24.554  < 2e-16 ***
## TaxRateRank  2.234e+04  5.600e+03   3.988 0.000128 ***
## MinTaxWage  -1.795e+02  4.447e+01  -4.037 0.000108 ***
## TFPerWages   3.638e+05  1.790e+05   2.032 0.044809 *
## TFWagesRank  1.012e+04  4.009e+03   2.524 0.013224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 231900 on 98 degrees of freedom
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.9691
## F-statistic: 471.8 on 7 and 98 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(LinearRegresssion_model2)
```




```r
acf(modback$residuals)
```



The model shows a strong relationship between StateRevenue and the predictors Year, Quarter, AvgTaxRate, TaxRateRank, MinTaxWage, TFPerWages, and TFWagesRank. It suggests that as time progresses (each year and within each year), state revenue increases, which could be due to economic growth, inflation, or changes in tax policy.

Higher average tax rates are associated with higher revenue, which is expected. The negative relationship with MinTaxWage might need further investigation, as it does not align with typical economic theories where higher minimum wages could lead to increased spending and thus higher revenue. It could be related to other economic activities or policies that are not captured in the model.

The trust fund-related variables (TFPerWages and TFWagesRank) indicate a positive relationship with state revenue, suggesting that better management or size of the trust fund relative to wages is beneficial for state revenue.

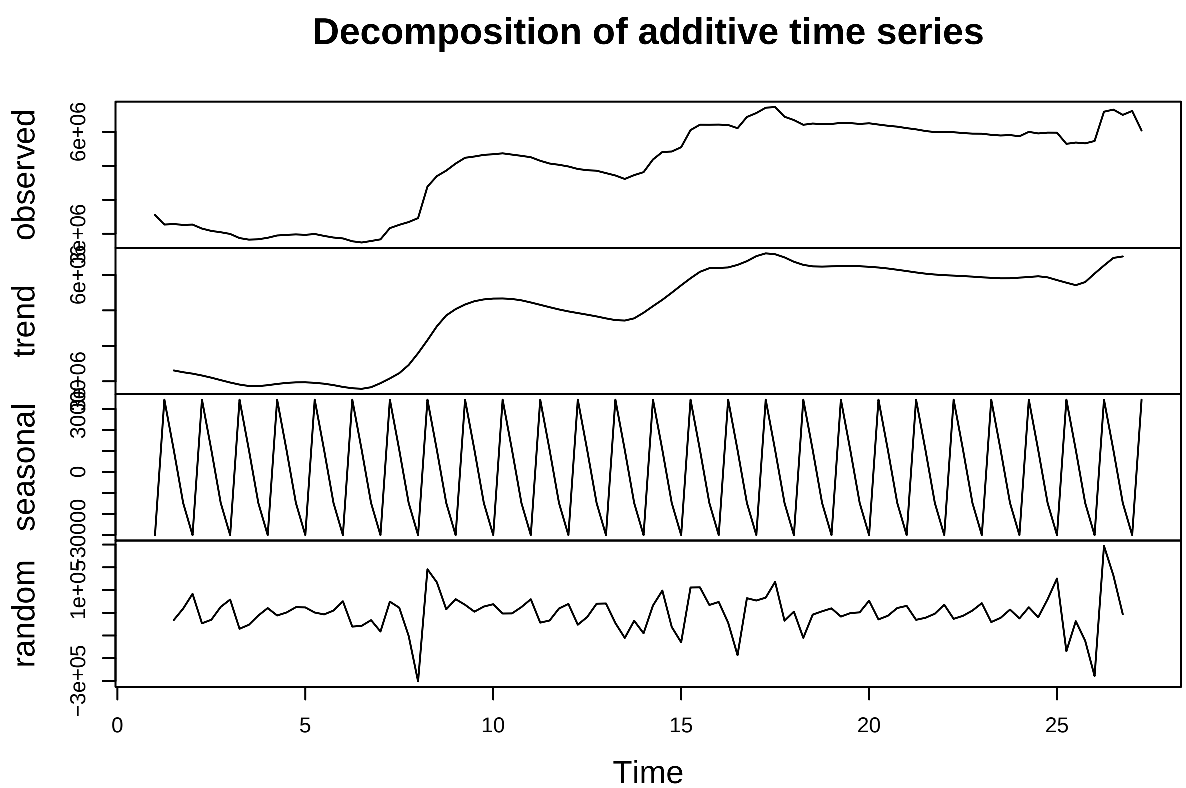
The model is statistically robust and provides valuable insights for policymakers and economists interested in the factors that influence state revenue. It is important to note that while the model has a high explanatory power, the causal relationships should be investigated further before making policy decisions.

```
#ARIMA Model
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.3.3
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

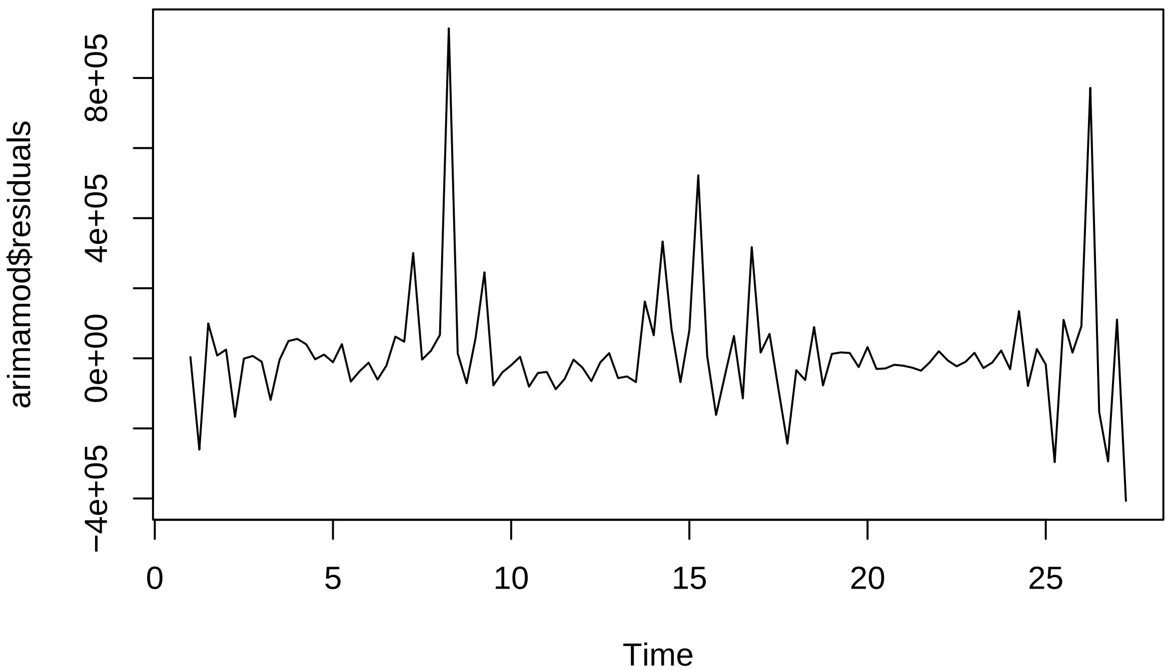
```r
y=ts(final_data$StateRevenue,frequency=4)
plot(decompose(y))
```



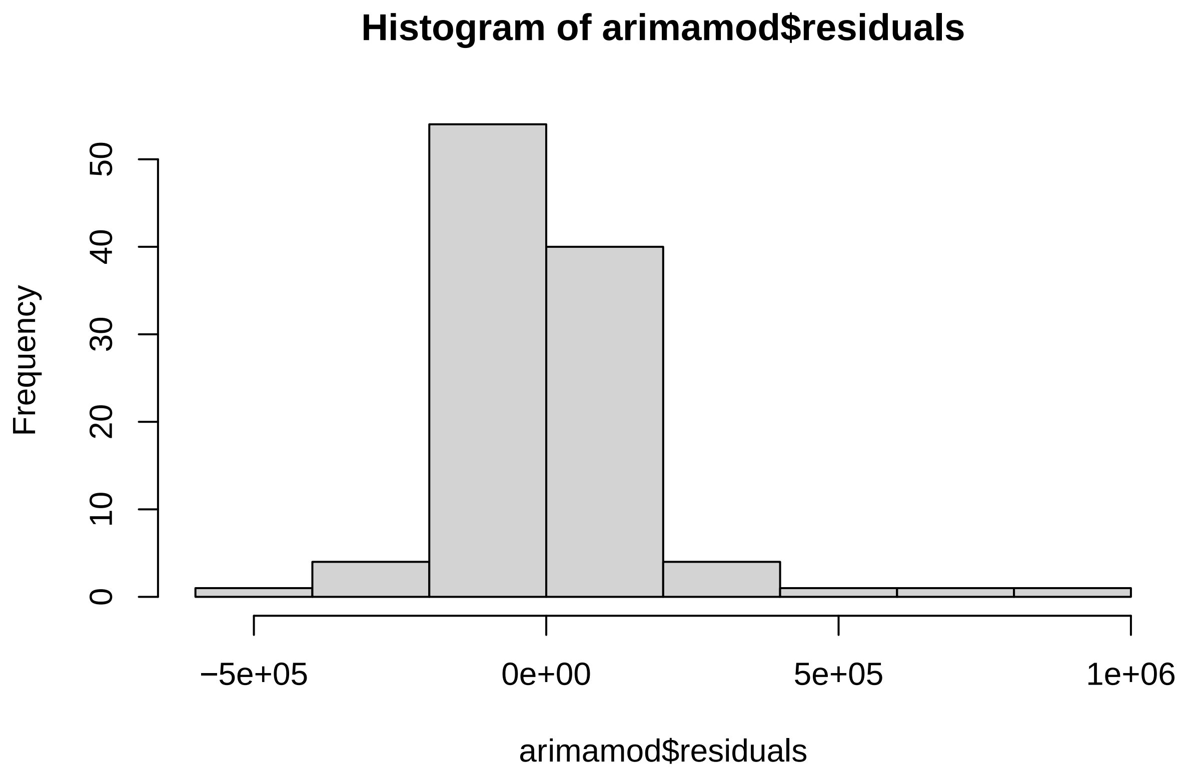

```r
arimamod<-auto.arima(ts(final_data$StateRevenue,frequency=4))
arimamod
```

```
## Series: ts(final_data$StateRevenue, frequency = 4)
## ARIMA(1,1,1)(1,0,0)[4]
##
## Coefficients:
##          ar1      ma1     sar1
##       0.7847  -0.5019  -0.2510
## s.e.  0.1316   0.1788   0.1101
##
## sigma^2 = 2.881e+10:  log likelihood = -1412.06
## AIC=2832.12   AICc=2832.52   BIC=2842.73
```

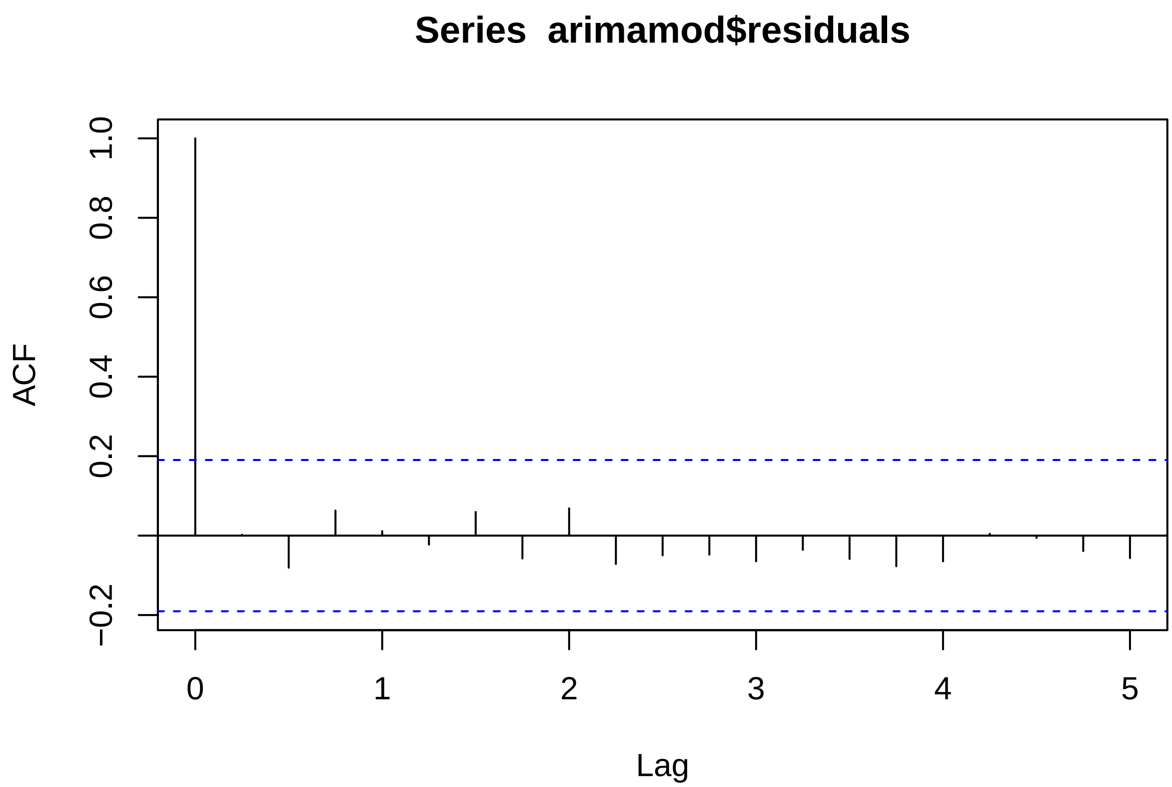
```r
#par(mfrow = c(2, 2))
plot(arimamod$residuals)
```




```r
hist(arimamod$residuals)
```

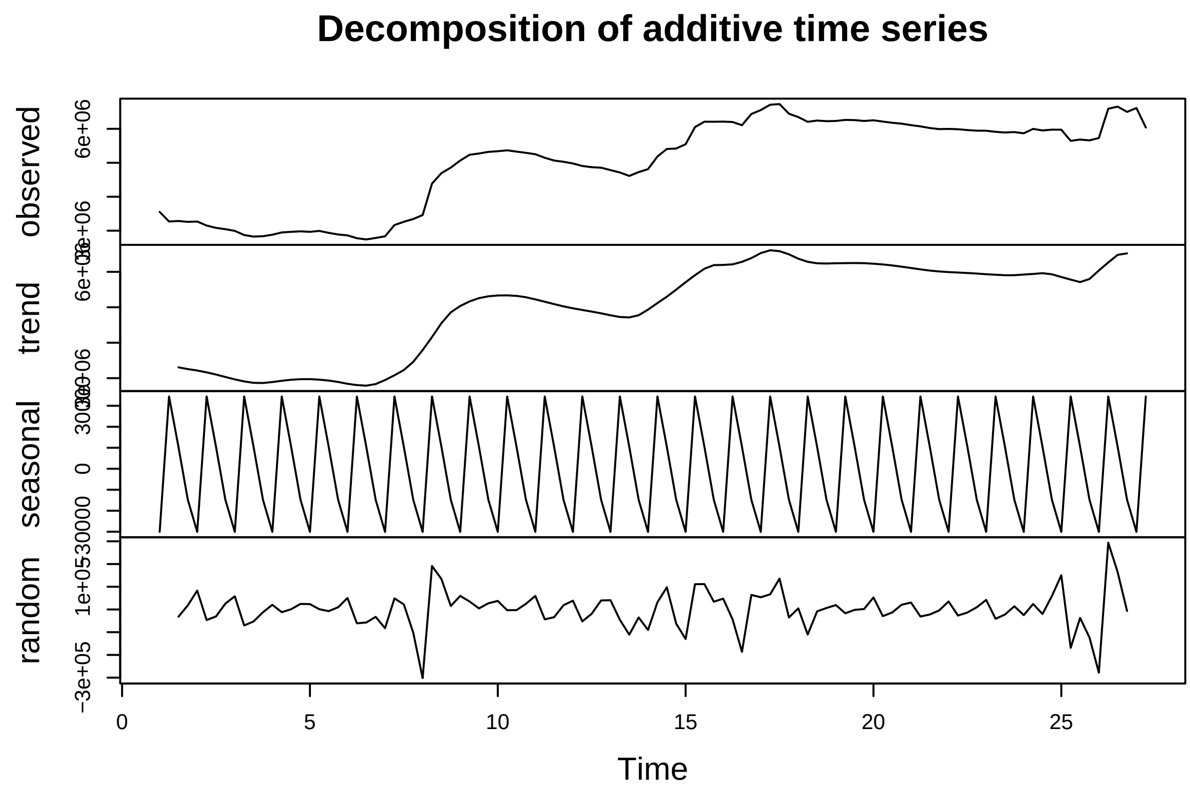



```
acf(arimamod$residuals)
```



The ARIMA model suggests that StateRevenue is significantly affected by its past values, both from the last term and from the last year's same quarter, though the seasonal effect is slightly negative. There is also an adjustment for fluctuations from the recent past. This model could potentially be used to forecast future revenue, but the large error variance indicates predictions might have substantial uncertainty. It would be good to compare this model's AIC and BIC with other models to choose the best one for forecasting.

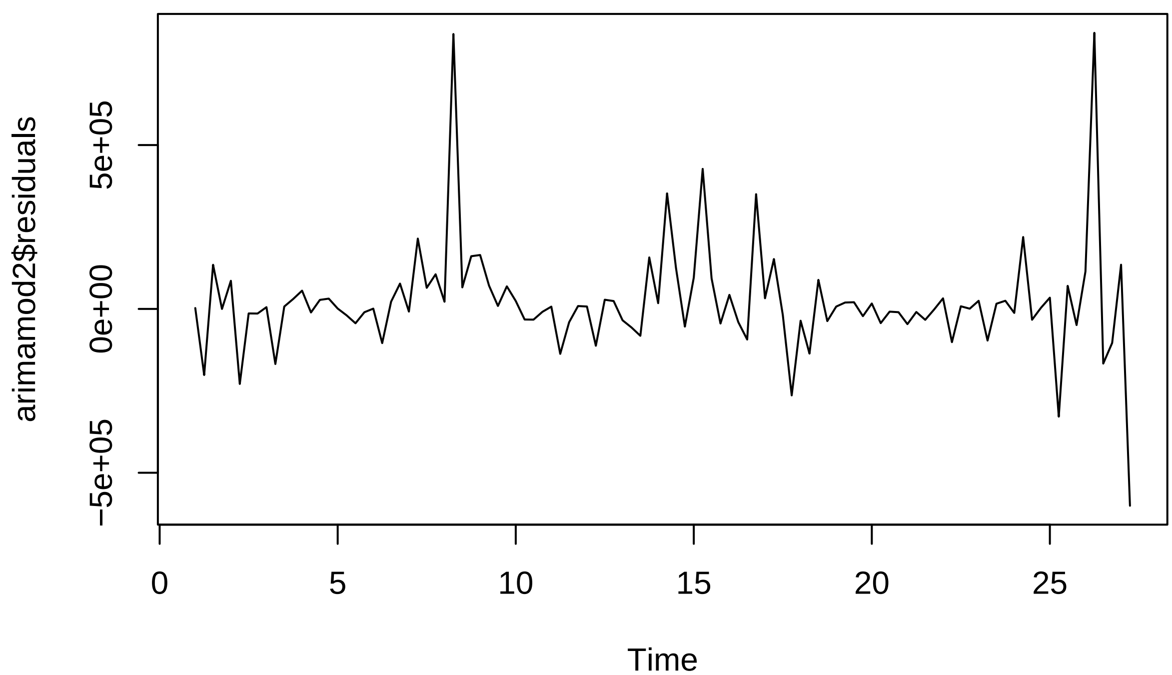
```
library(forecast)
X<-as.matrix(final_data[,c("AvgTaxRate", "TrustFund"  , "TFWagesRank" , "Interest" , "HighCostMultiple"
y=ts(final_data$StateRevenue,frequency=4)
plot(decompose(y))
```



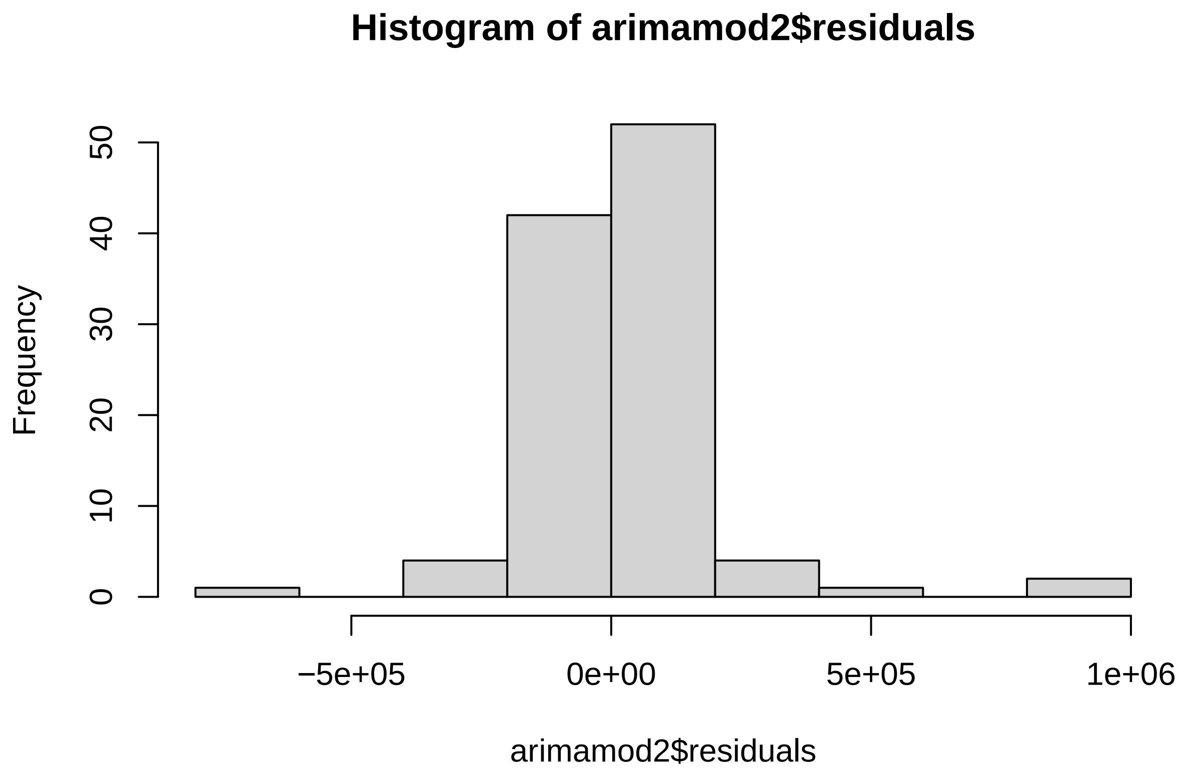

```
arimamod2<-Arima(y,order = c(0,1,1),xreg=X)
arimamod2
```

```
## Series: y
## Regression with ARIMA(0,1,1) errors
##
## Coefficients:
##          ma1  AvgTaxRate  TrustFund  TFWagesRank  Interest  HighCostMultiple
##       0.2727    671762.7     0.0187     4030.341    1.5670          153214.3
## s.e.  0.1424    772678.8     0.0245     6625.110    2.5258          239455.5
##
## sigma^2 = 3.049e+10:  log likelihood = -1413.32
## AIC=2840.63   AICc=2841.79   BIC=2859.21
```

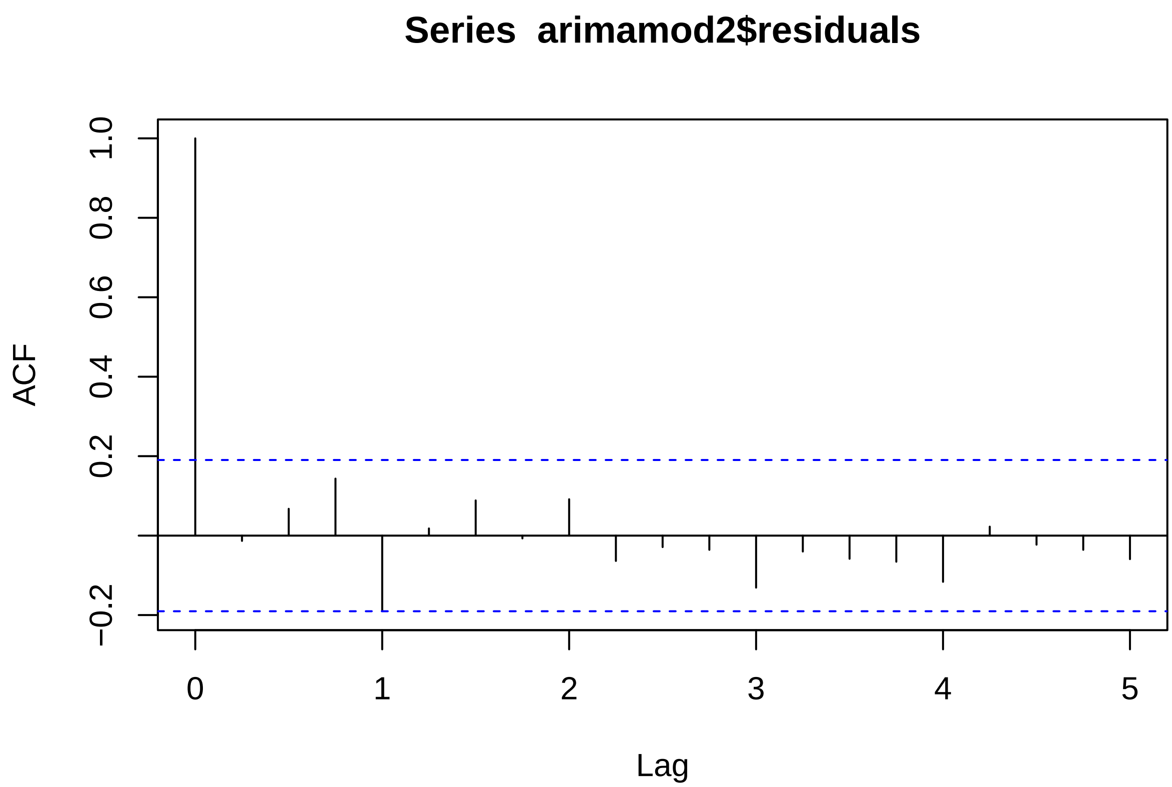
```
#par(mfrow = c(2, 2))
plot(arimamod2$residuals)
```




```r
hist(arimamod2$residuals)
```




```
acf(arimamod2$residuals)
```



This model suggests that StateRevenue is significantly influenced by its previous values (as indicated by the ARIMA component), and by the factors included in the regression part, such as AvgTaxRate, TFWagesRank, and HighCostMultiple. The average tax rate has the most substantial impact on revenue, with each increase leading to a significant increase in revenue. The error terms also tell us that there's a slight tendency for errors to follow a pattern over time, which is being corrected by the model.

While the model seems to fit the data well (as suggested by the statistical significance of the coefficients), it does exhibit a high variability in its predictions (sigma^2 value). The model's usefulness would be in its predictive power, which should be evaluated against actual outcomes to determine its accuracy. When choosing the best model for predicting StateRevenue, one should consider both the AIC/BIC values and the context in which the model will be applied.

```
#install.packages("car")
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

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

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
vif_model <- vif(lm(StateRevenue ~ AvgTaxRate + Year + Quarter + TrustFund + TFWagesRank + Interest + Hi
print(vif_model)
```

```
##       AvgTaxRate              Year           Quarter         TrustFund
```

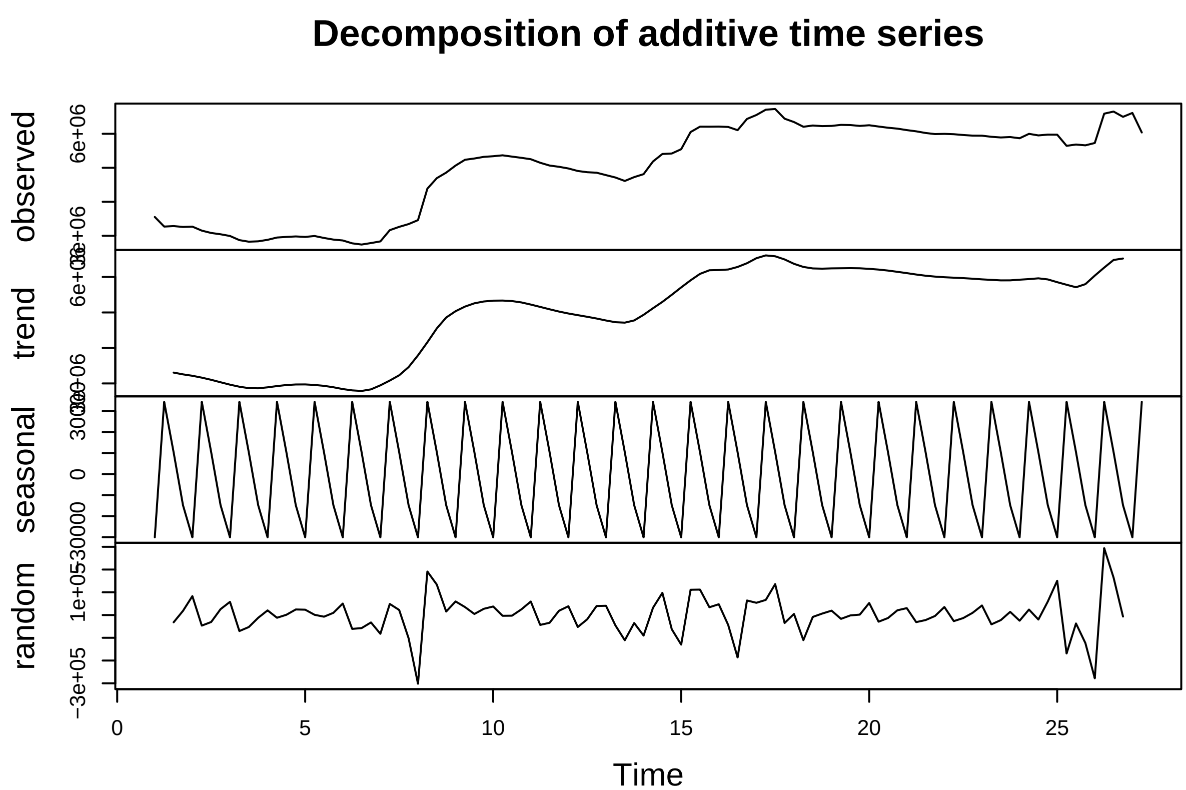
```
##         1.260259         1.503159         1.012298         2.899734
##      TFWagesRank         Interest HighCostMultiple
##         1.536068         2.791886         1.143677
```

```r
library(forecast)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
```

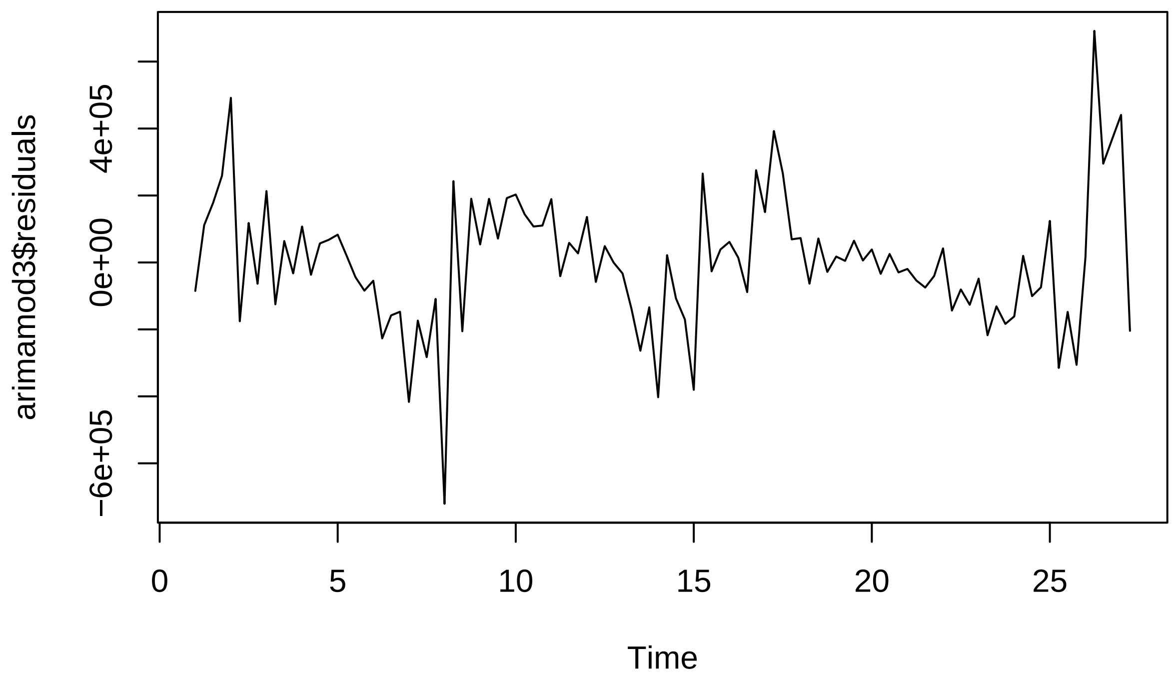
```r
# Prepare exogenous variables matrix
X <- as.matrix(final_data[, c("AvgTaxRate", "Year", "Quarter", "Interest", "HighCostMultiple")])

# Center and scale predictors
X_scaled <- scale(X)

# Create time series object
y <- ts(final_data$StateRevenue, frequency = 4)
plot(decompose(y))
```



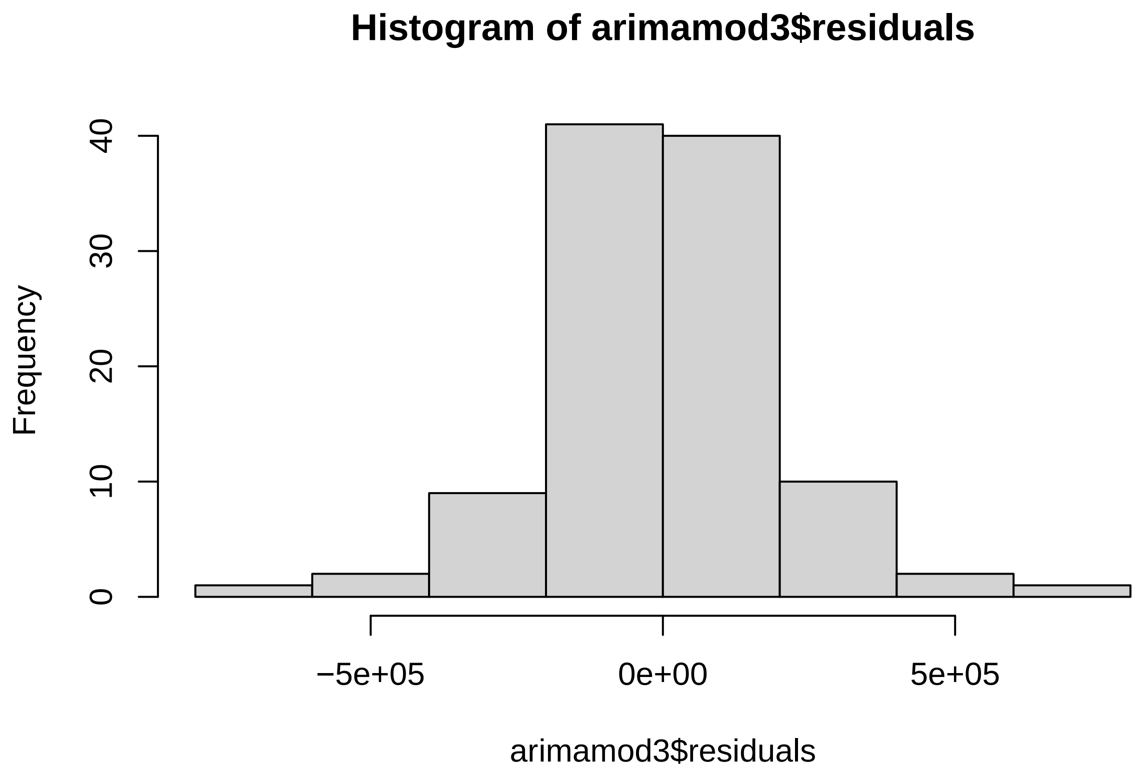

```r
# Fit ARIMA model
arimamod3 <- Arima(y, order = c(0, 0, 1), xreg = X_scaled)
arimamod3
```

```
## Series: y
## Regression with ARIMA(0,0,1) errors
##
## Coefficients:
##          ma1   intercept  AvgTaxRate        Year   Quarter  Interest
##       0.6473  5008054.16   609922.16  1341622.59  63321.92  43601.29
## s.e.  0.0606    31190.29    32768.93    35838.69  13556.37  28104.16
##       HighCostMultiple
##               31750.68
## s.e.          19293.30
##
## sigma^2 = 4.098e+10:  log likelihood = -1442.18
## AIC=2900.36   AICc=2901.84   BIC=2921.66
```

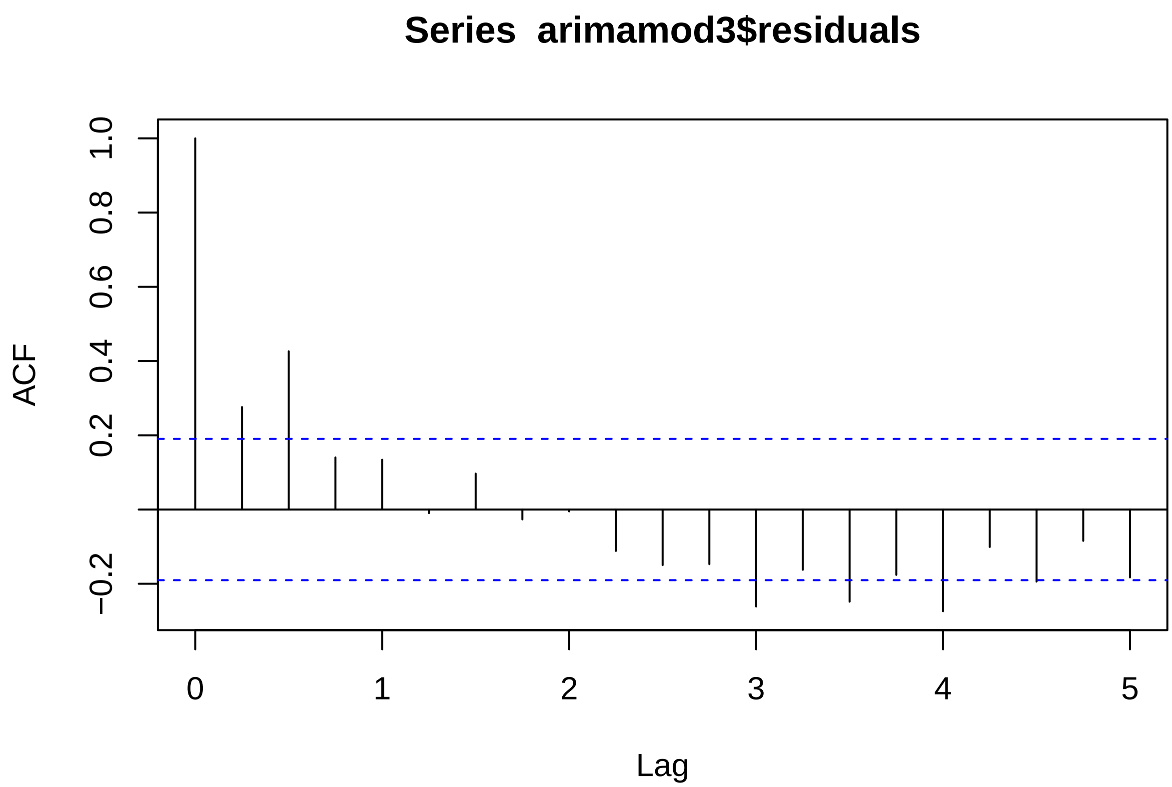
```r
#par(mfrow = c(2, 2))
plot(arimamod3$residuals)
```




```r
hist(arimamod3$residuals)
```




```
acf(arimamod3$residuals)
```



This model indicates that StateRevenue is affected by several factors, including the average tax rate, the year, and the quarter. The tax rate has a particularly strong influence, which makes sense economically, as higher taxes generally mean higher revenue. The positive coefficients for Year and Quarter suggest an overall increase in revenue over time and possible seasonal effects within each year.

The ARIMA component with the moving average term suggests there's a pattern in the errors from one period to the next that the model is accounting for. The relatively high sigma^2 value points to a considerable amount of unexplained variability, which could imply the presence of other influential factors not included in the model or inherently unpredictable fluctuations in revenue.

While the model may be useful for understanding and forecasting revenue to some extent, its accuracy and predictive power would ideally be assessed against actual revenue figures and compared with other models using AIC and BIC values before making definitive conclusions.

```r
library(readxl)

# Read the data from the Excel file
data <- read_excel("C:/Users/Suma2/Downloads/Applied stats_projrct/cawlifornis.xlsx")

# Set the seed for reproducibility
set.seed(123)

# Shuffle the data
data <- data[sample(nrow(data)), ]

# Splitting the data
test_data <- data[1:10, ]  # Test data with 10 records
train_data <- data[11:nrow(data), ]  # Training data with the rest
```

```r
# You can now fit models on your train_data and evaluate them on test_data.

# Assuming you have fitted models and calculated AIC for each, here's how you could compare them:
# (Note: Replace model_aic_1, model_aic_2, etc., with actual AIC values from your models)

model_aic_1 <- AIC(LinearRegresssion_model1)  # Replace model1 with your actual model
model_aic_2 <- AIC(LinearRegresssion_model2)
model_aic_3 <- AIC(modback)
model_aic_4 <- AIC(arimamod)
model_aic_5 <- AIC(arimamod2)
model_aic_6 <- AIC(arimamod3)

# Collecting all AICs for comparison
aic_values1 <- c(model_aic_1, model_aic_2,model_aic_3)
aic_values2 <- c(model_aic_4, model_aic_5, model_aic_6)


# Finding the model with the minimum AIC
min_aic <- min(aic_values1)
min_aic1 <- min(aic_values2)
best_model_index <- which(aic_values1 == min_aic)
best_model_index1 <- which(aic_values2 == min_aic1)

# Print the best model index

print(paste("AIC of Linear Regression model 1 is ", model_aic_1))
```

```
## [1] "AIC of Linear Regression model 1 is  2099.83925897627"
```

```r
print(paste("AIC of Linear Regression model 2 is ", model_aic_2))
```

```
## [1] "AIC of Linear Regression model 2 is  2069.51568325069"
```

```r
print(paste("AIC of Linear Regression model 3 is ", model_aic_3))
```

```
## [1] "AIC of Linear Regression model 3 is  2929.56848957018"
```

```r
print(paste("AIC of Arima model 1 is ", model_aic_4))
```

```
## [1] "AIC of Arima model 1 is  2832.11793458063"
```

```r
print(paste("AIC of Arima model 2 is ", model_aic_5))
```

```
## [1] "AIC of Arima model 2 is  2840.63263739961"
```

```r
print(paste("AIC of Arima model 3 is ", model_aic_6))
```

```
## [1] "AIC of Arima model 3 is  2900.35677876141"
```

```r
print(paste("The best Linear model is ", best_model_index, "with an AIC of", min_aic))
```

```
## [1] "The best Linear model is  2 with an AIC of 2069.51568325069"
```

```r
print(paste("The best Arima  Model is Model ", best_model_index1, "with an AIC of", min_aic1))
```

```
## [1] "The best Arima  Model is Model  1 with an AIC of 2832.11793458063"
```

So,The linear Model2 is the best among the two linear models.It provides a more detailed view by considering the year-over-year changes and potential seasonal effects within the year, although the latter did not show a significant effect. The inclusion of time variables has slightly improved the model's explanatory power, making it a potentially more accurate tool for forecasting New Jersey's state revenue

The Basic ARIMA model is a relatively simple time series model that is useful for forecasting state revenue based on its past changes and the relationship of past errors to the current prediction.The Basic Arima is the best among the three arima models With less auto correlation value