MTH 325, Final Exam, At Home, Spring 2024 Name KEY

Instructions: This exam is in two parts: Part I is to be completed partly at home using the materials posted in the course for the at-home portion and you will answer questions about that work during the in-class portion of the exam; Part II is to be completed entirely in class. You may not use cell phones, and you may only access internet resources you are specifically directed to use.

At home, prepare for questions in Part I using R. Open the data file entitled **325final_data.xlsx** posted in Blackboard. Complete the calculations noted below. You will be asked for additional analysis and interpretation of this data in the in-class portion of the test. Print out the results of your analysis and code, and bring the pages with you to the exam. You will submit all this work along with the in-class exam.

Use the data on motels to complete the following tasks. Sheet 1 has data on 990 purchases from a store. Sheet 2 has 10 additional purchases from the same store.

1. Import the data in the file into R and remove the Person column (it is not a variable). Separate your data into two dataframes. One for the Sheet 1 (training) and one for Sheet 2 (test data).
2. Conduct a two-way ANOVA test (with interaction) for the variables Age (category) and Salary (category) to predict Amount Spent. Create appropriate diagnostic plots for your model. Be prepared to describe your hypothesis tests and their outcomes.
3. Recreate the same model with the glm() function. Be prepared to discuss how the general linear model differs from the ANOVA model.
4. Convert all your categorical variables to dummy variables. Let your defaults be Young (Age), Female (Gender), ~~No~~ Rent (Home), No (Married), and Low Salary (Salary Category).
5. Create a correlation table of the variables. Make a correlation plot (type is of your choice), or a pairplot.
6. Create a multiple variable model of Amount Spent using all remaining available variables. Use appropriate automated selection techniques. Compare the result to manual backward selection. In your backward selection, stop only when all the coefficients are significant at the 0.05 level.
7. Construct diagnostic plots for your machine selected model and your manually selected model (these may be the same). Identify any potential problems with model assumptions, outliers and influential points.
8. Using the data on Sheet 2, predict the Amount Spent for the remaining customers. Compare the results to the provided Amount Spent values. Calculate your error.

To complete the calculations below, use the time series Seatbelts.

9. Create a new column in the dataset (or a separate vector) that represents the ratio of front-end crashes to rear-end crashes. Construct appropriate one-variable numerical plots to describe the overall data set.

10. Create a plot of the new time series. Perform seasonal decomposition and plot the resulting graph.

11. Create an ACF and PACF graph for the time series.

12. Construct an ARIMA model. Plot the model against the original time series. You may need to experiment with settings to select the best combination of p, d, and q.

MTH 325, Final Exam, Spring 2024 Name KEY

Instructions: Answer each question thoroughly. For questions in Part 1, use the work you did at home to answer the questions. Be sure to answer each part of each question. In Part 2, report exact answers unless directed to round.

Part I:

Use the work you did at home to answer these questions about spending.

1. Based on your correlation table or correlation plot, identify the variable that has the highest negative correlation with Amount Spent. What is the (approximate) correlation value?

Male it's very small negative but it is the most negative

2. Based on your ANOVA model, should interactions be included in your model or not?

they should not. interaction p-value is 0.7109

3. Do the residuals from your ANOVA or general linear model appear to be normally distributed?

they do not appear normal (nor does Amount Spent)

4. After converting the categorical variables to dummy variables, which two variables appear to have the highest correlation (positive or negative)?

Married w/ Huge Salary around 0.588

5. After performing backward selection, what is the R^2 value of your resulting model?

0.048
or 4.8%

6. Write the equation you obtained from your backward selection process for predicting operating expenses. Be sure to clearly indicate what each variable in the equation represents.

Amt Spent = 279.11 * Married + 662.99

7. Describe how your other model selection methods differed (or were similar to) the results obtained from the backward selection process.

machine selected model kept Married and also Medium Salary, High Salary and Huge Salary and gained half % of R^2

Medium Salary fails significance

8. What percentage of the variability in Amount Spent can be explained by the relationship to the other model variables?

4.8% or 5.5%

backward machine stepwise

9. Answer this question and the remaining questions in Part 1 using the backward selection model you found by hand. Do your diagnostic plots suggest any outliers or model problems? Explain.

Amount Spent & residuals do not appear normal

producing several outliers

the model has very little predictive power

10. How do your predictions for the 10 extra people? How does your residual error (RMSE) differ from the model residual error?

The RMSE is 612.9984

which is basically identical to the model residual error

which is very large given size of the predictive means

11. Interpret the meaning of the Married coefficient in the context of the problem.

a married person will spend, on average
an additional $279.11

12. If you needed to build a model of Amount Spent with two variables, what would they be, and why?

according to best subset regression use Middle & Older (Age Category)

Use the work you did at home to answer these questions about the time series model.

13. Does the model appear approximately stationary or does there appear to be a trend? Consider any boxplots or histograms here, as well as any time series plots or decompositions you may have done.

yes, until the law changes, then there is a dip
the trend is pretty flat until then

14. Based on your PACF graph, how many lags should be included in your time series model? Why?

1 only one is above the significance line
until half a year out

15. What settings did you use for your ARIMA model? Why? What diagnostics did you use to select these settings?

used ARIMA (1,1,5)

produced the lowest AIC (5.9)

and consistent w/ ACF and PACF graphs

16. Write the equation of your final time series model.

$$\hat{y}_t = y_{t-1} + 0.2623\,(y_{t-1} - y_{t-2}) - 0.6308\,e_{t-1} + 0.089\,e_{t-2} - 0.0650\,e_{t-3}$$

$$- 0.0689\,e_{t-4} - 0.0855\,e_{t-5}$$

17. What is the AIC of your final model? How good does the model appear to fit?

5.9

based on graph and other metrics
this is a good fit

Part II:

18. Recall that $Cov(X,Y) = E(XY) - E(X)E(Y)$. For the probability density function $f(x,y) = \frac{1}{2}x^2(y+1), y \in [0,1], x \in [0,1]$, find the covariance.

$$E(X) = \int_0^1 \int_0^1 \tfrac{1}{2}\, x \cdot x^2 (y+1)\, dy\, dx = \frac{1}{2} \cdot \frac{3}{8} = \frac{3}{16}$$

$$E(Y) = \int_0^1 \int_0^1 \tfrac{1}{2}\, y \cdot x^2 (y+1)\, dy\, dx = \frac{1}{2} \cdot \frac{5}{18} = \frac{5}{36}$$

$$E(XY) = \int_0^1 \int_0^1 xy \cdot x^2 (y+1)\, dy\, dx = \frac{1}{2} \cdot \frac{5}{24} = \frac{5}{48}$$

$$E(XY) - E(X)E(Y) = \frac{5}{48} - \frac{3}{16} \cdot \frac{5}{36} =$$

$$\frac{5}{48} - \frac{5}{192} = \frac{5}{64} = 0.078125$$

19. Consider the small data set $\{(3,2),(5,4),(9,9)\}$. Find the value of the regression coefficients for $y = \beta_0 + \beta_1 x$, using the normal equation $(A^T A)^{-1} A^T Y = B$. Write the coefficients you find in the equation.

$$A = \begin{bmatrix} 3 & 1 \\ 5 & 1 \\ 9 & 1 \end{bmatrix} \quad A^T A = \begin{bmatrix} 3 & 5 & 9 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 5 & 1 \\ 9 & 1 \end{bmatrix} = \begin{bmatrix} 115 & 17 \\ 17 & 3 \end{bmatrix} \quad A^T Y = \begin{bmatrix} 3 & 5 & 9 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 9 \end{bmatrix} = \begin{bmatrix} 107 \\ 15 \end{bmatrix}$$

$$B = \begin{bmatrix} 33/28 \\ -47/28 \end{bmatrix} \approx \begin{bmatrix} 1.17857 \\ -1.67857 \end{bmatrix}$$

$$y = 1.17857x - 1.67857$$

20. Suppose that a classification model (such as logistic regression) produced the following confusion matrix. Calculate the accuracy and discuss whether the model result reveals any potential problems.

	Yes	No
Yes	641	11
No	24	2

$$\frac{643}{678} = 94.8\%$$

this shows masking
more nos are miscategorized than correctly categorized

the no category is also very small compared to yes

21. Describe clustering (in machine learning) and give an example of a machine learning algorithm that implements this learning method. Is this method an example of supervised, unsupervised or semi-supervised learning.

K-means (answers may vary)

Clustering algorithms try to spot relationships in points

Clustering is unsupervised (or can be semi-supervised)

which means labels are not used to help identify groupings

22. Describe how Gaussian process regression works in general terms.

generally gaussian processes find a weighted mean (by distance) of nearby points to estimate a mean model

23. What are some reasons it might be beneficial to use a non-parametric nonlinear model for a regression problem rather than a parametric non-linear model?

may provide error estimates, may be more flexible than parametri polynomial models, makes fewer assumptions about data

24. What is one reason you might get an error from the decompose() function applied to a time series?

if there is no decernible seasonal pattern.

25. Explain why autocorrelation prevents us from using traditional regression to model some time series data.

traditional regression assumes errors are independent while time series errors may not be.

26. Why are irregular time series so much more difficult to work with than regular time series? Describe some methods we can use to make irregular time series more regular.

most methods for dealing w/ time series are based on regular ones so these methods are not available w/o interpolation or resampling which can ~~_~~ result in adding bias or losing data.
regression makes other assumptions that may not apply

27. How do we use the AUC (of an ROC curve) as a diagnostic for a classification model?

helps us choose optimal model
and is related to accuracy of model
higher is better, closer to 50% is a coin toss

28. Describe the difference between LASSO and Ridge regression. Explain how the penalty helps to address the bias-variance trade-off.

the difference is related to the penalty. one uses a first order penalty while the other uses a square penalty
The penalty helps reduce the effect of variables that don't contribute to reducing variance to avoid overfitting

29. How does Spearman's correlation differ from Pearson correlation?

Spearman predicts increase or decrease but not linearity

30. What are some potential advantages and disadvantages of using variable transformations in a model?

They may help improve model fit or fix heteroscedacity
but they could also introduce problems (such as heteroscedacity if it was not there before)

31. What is the difference between an outlier and an influential point in a regression model?

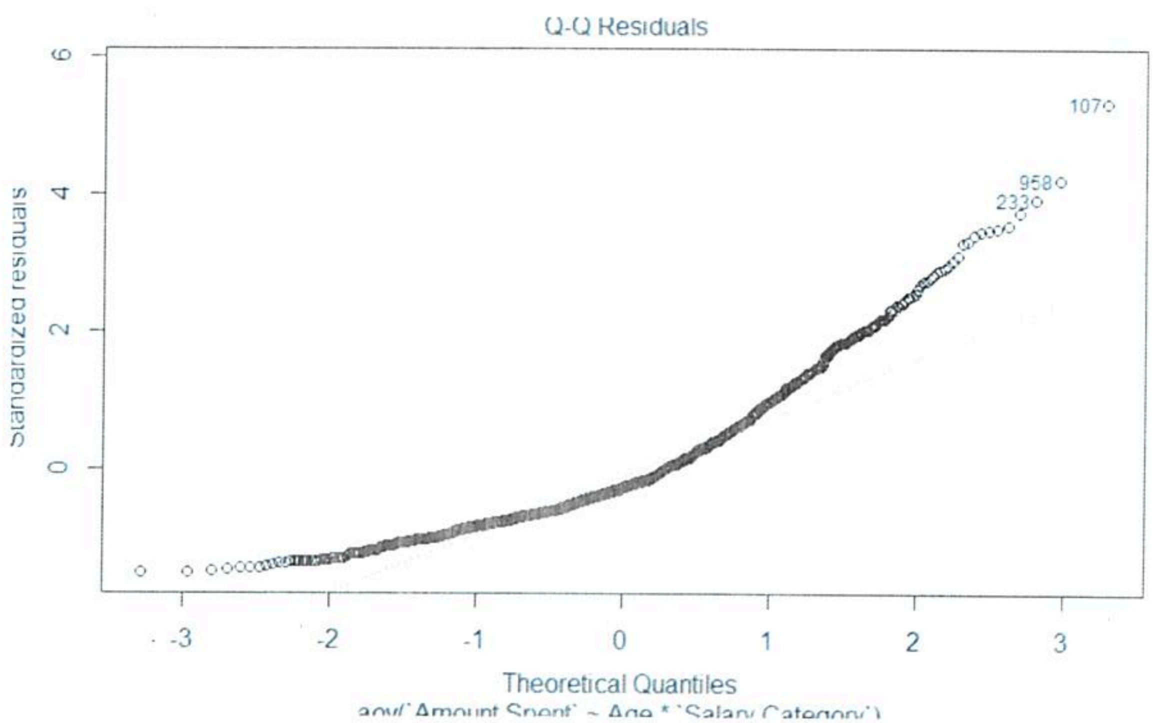
a point may be one or both. an influential point has a large impact on model coeff while an outlier has a large residual.

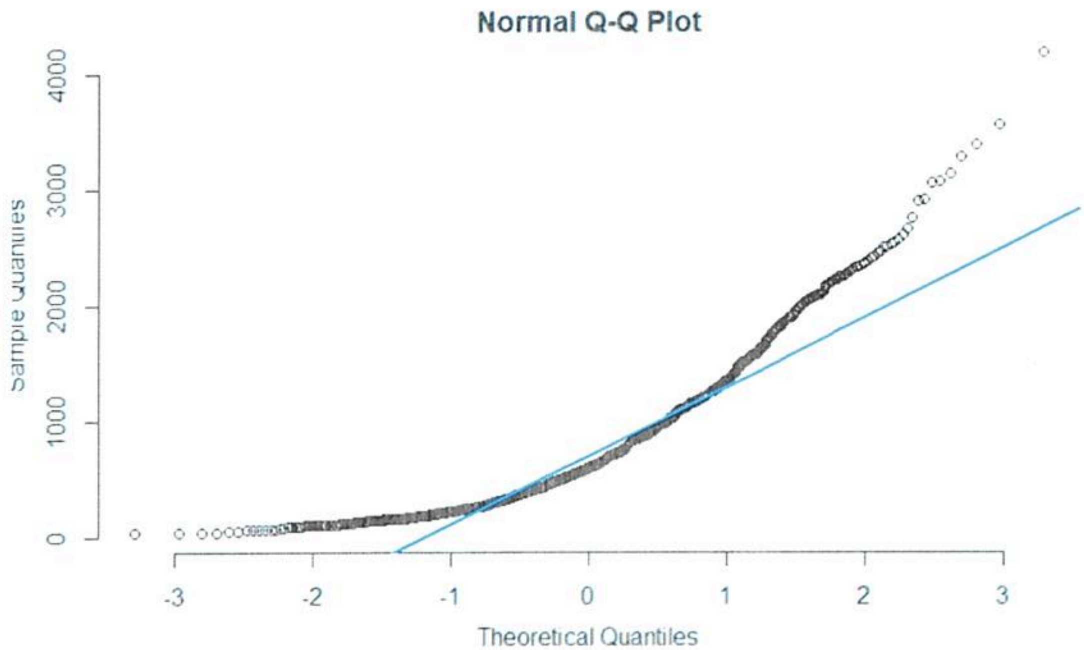
MTH 325 Final Exam at-home analysis

```
                         Df    Sum Sq Mean Sq F value    Pr(>F)    
Age                       2   6460183 3230092   8.548 0.000209 ***
`Salary Category`         3  12677800 4225933  11.184 3.21e-07 ***
Age:`Salary Category`     6   1415516  235919   0.624 0.710920    
Residuals               978 369547284  377860                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                        (Intercept)                                        AgeOlder 
                         790.555556                                      142.324444 
                           AgeYoung                `Salary Category`Huge Salary 
                         140.325397                                      192.054614 
        `Salary Category`Low Salary              `Salary Category`Medium Salary 
                        -416.555556                                       -4.227197 
AgeOlder:`Salary Category`Huge Salary   AgeYoung:`Salary Category`Huge Salary 
                        -126.193873                                      -41.735566 
 AgeOlder:`Salary Category`Low Salary    AgeYoung:`Salary Category`Low Salary 
                          52.056508                                      111.387914 
AgeOlder:`Salary Category`Medium Salary AgeYoung:`Salary Category`Medium Salary 
                        -288.442276                                     -203.875977 
```

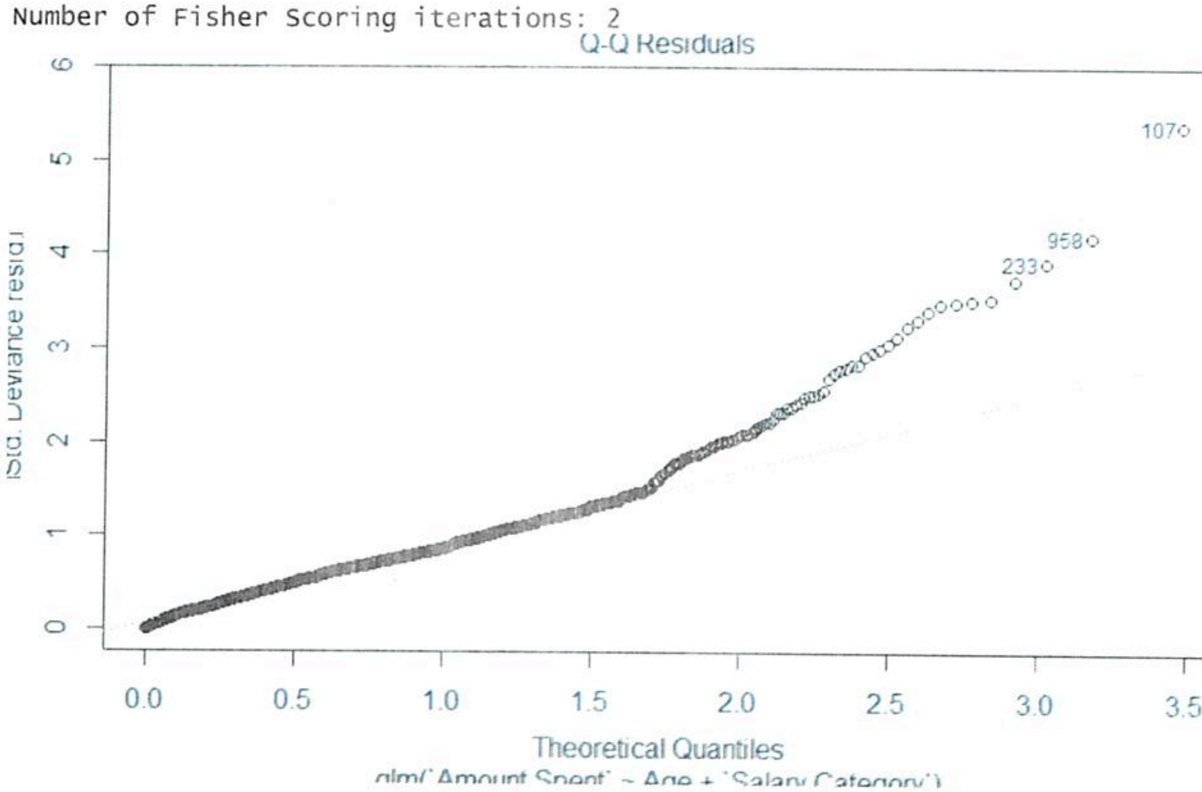
```
                     Df    Sum Sq Mean Sq F value    Pr(>F)    
Age                   2   6460183 3230092   8.568 0.000205 ***
`Salary Category`     3  12677800 4225933  11.210 3.09e-07 ***
Residuals           984 370962800  376995                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

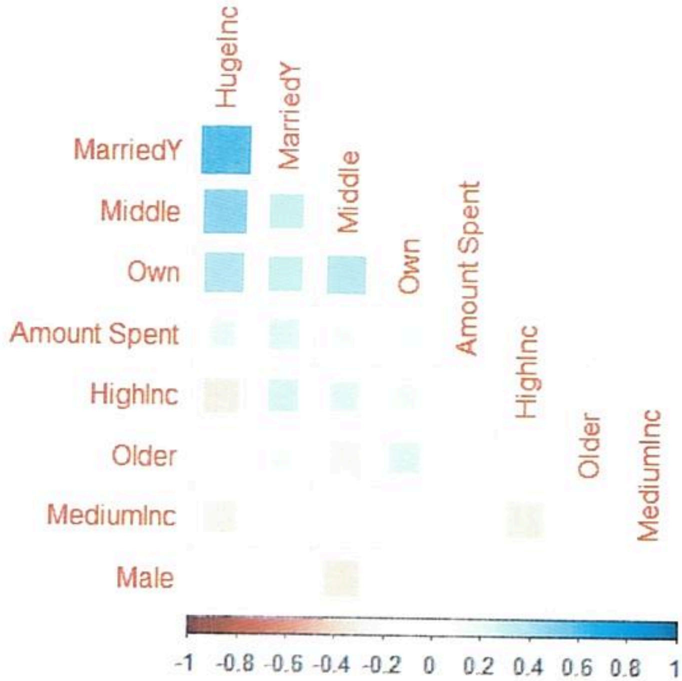


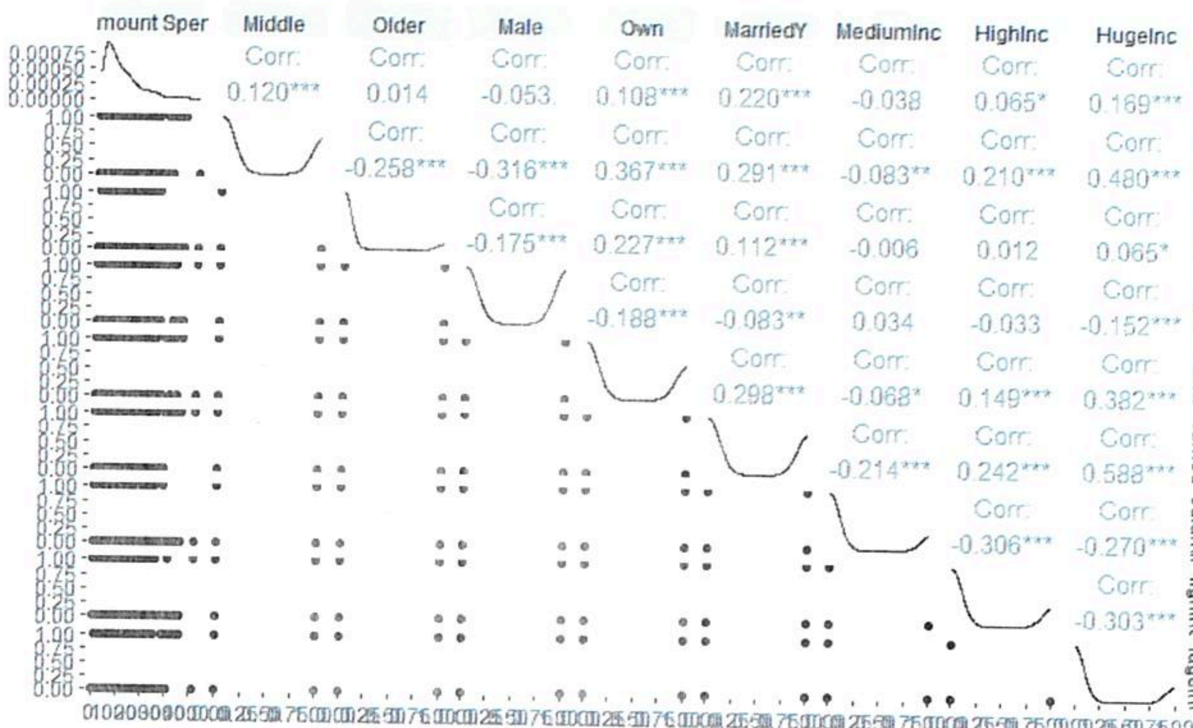



```
Call:
glm(formula = `Amount Spent` ~ Age * `Salary Category`, data = data1)

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              790.556     51.225  15.433  < 2e-16
***
AgeOlder                                 142.324    133.186   1.069  0.28551
AgeYoung                                 140.325     84.394   1.663  0.09669
.
`Salary Category`Huge Salary             192.055     68.984   2.784  0.00547
**
`Salary Category`Low Salary             -416.556    437.669  -0.952  0.34145
`Salary Category`Medium Salary            -4.227     90.905  -0.047  0.96292
AgeOlder:`Salary Category`Huge Salary   -126.194    184.033  -0.686  0.49306
AgeYoung:`Salary Category`Huge Salary    -41.736    291.255  -0.143  0.88609
AgeOlder:`Salary Category`Low Salary      52.057    473.985   0.110  0.91257
AgeYoung:`Salary Category`Low Salary     111.388    444.232   0.251  0.80207
AgeOlder:`Salary Category`Medium Salary -288.442    208.004  -1.387  0.16584
AgeYoung:`Salary Category`Medium Salary -203.876    125.543  -1.624  0.10471
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 377860.2)

    Null deviance: 390100784  on 989  degrees of freedom
Residual deviance: 369547284  on 978  degrees of freedom
AIC: 15537

                          (Intercept)                               AgeOlder
                           790.555556                             142.324444
                             AgeYoung          `Salary Category`Huge Salary
                           140.325397                             192.054614
          `Salary Category`Low Salary        `Salary Category`Medium Salary
                          -416.555556                              -4.227197
AgeOlder:`Salary Category`Huge Salary  AgeYoung:`Salary Category`Huge Salary
                          -126.193873                             -41.735566
AgeOlder:`Salary Category`Low Salary    AgeYoung:`Salary Category`Low Salary
                            52.056508                             111.387914
```

```
AgeOlder:`Salary Category`Medium Salary AgeYoung:`Salary Category`Medium Salary
                             -288.442276                                   -203.875977

Number of Fisher Scoring iterations: 2

Call:
glm(formula = `Amount Spent` ~ Age + `Salary Category`, data = data1)

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                          829.02      44.82  18.498  < 2e-16 ***
AgeOlder                              22.28      72.80   0.306   0.7597
AgeYoung                              60.20      57.40   1.049   0.2945
`Salary Category`Huge Salary         153.71      59.85   2.568   0.0104 *
`Salary Category`Low Salary         -266.01      61.60  -4.318 1.73e-05 ***
`Salary Category`Medium Salary      -131.33      59.06  -2.224   0.0264 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 376994.7)

    Null deviance: 390100784  on 989  degrees of freedom
Residual deviance: 370962800  on 984  degrees of freedom
AIC: 15529

Number of Fisher Scoring iterations: 2
```







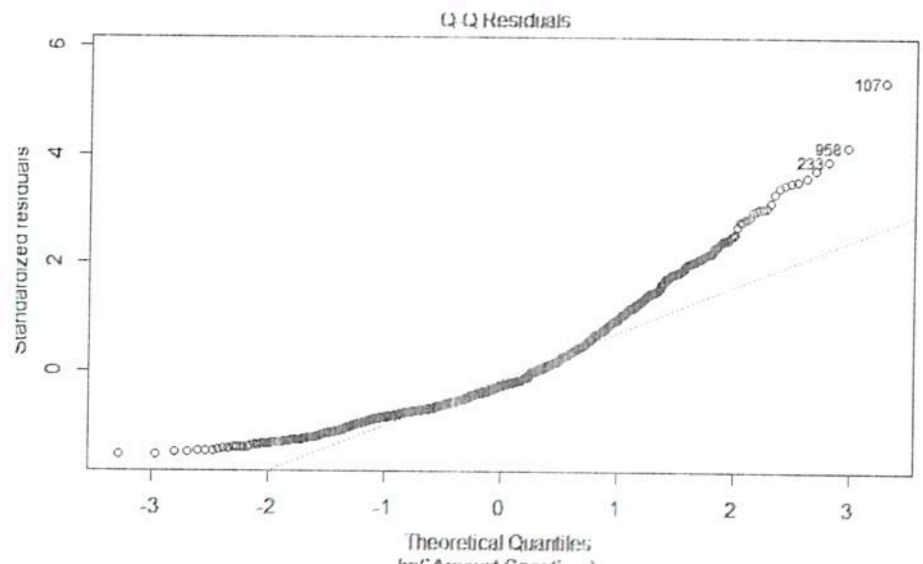

```
Call:
lm(formula = `Amount Spent` ~ ., data = data1)

Residuals:
   Min     1Q Median     3Q    Max 
-943.0 -429.9 -156.7  271.6 3270.7 

Coefficients:
```

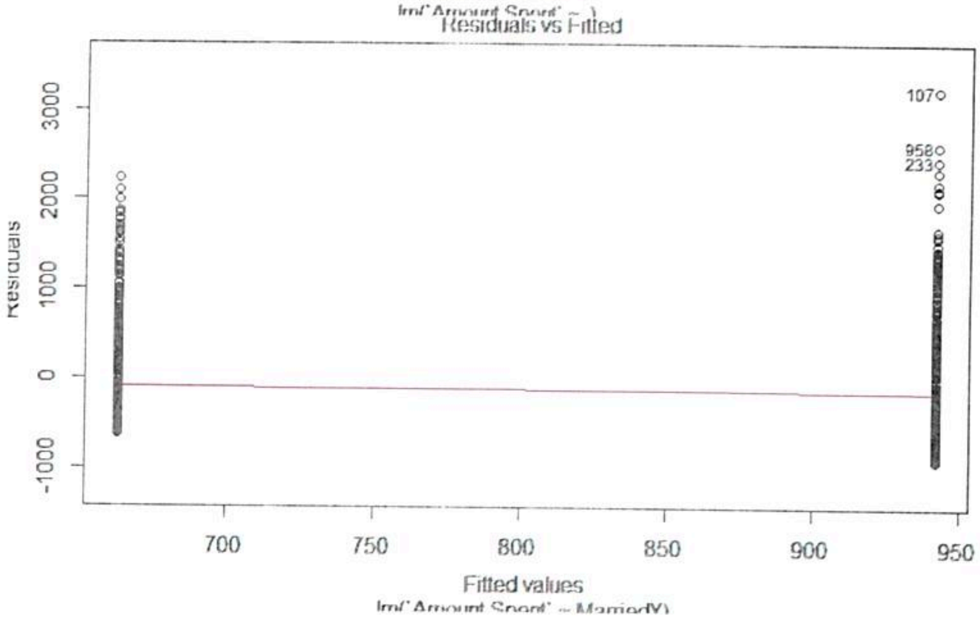
```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      638.66      44.49  14.356   <2e-16 ***
Middle           -23.70      66.64  -0.356   0.7222
Older            -50.03      80.10  -0.625   0.5324
Male             -34.86      43.33  -0.804   0.4214
Own               17.84      48.58   0.367   0.7136
MarriedY         164.87      64.56   2.554   0.0108 *
MediumInc         84.92      61.81   1.374   0.1698
HighInc          134.60      83.35   1.615   0.1066
HugeInc          211.37     114.96   1.839   0.0663 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 612.6 on 981 degrees of freedom
Multiple R-squared:  0.05622,	Adjusted R-squared:  0.04852
F-statistic: 7.305 on 8 and 981 DF,  p-value: 1.853e-09

Backward selection model:
Call:
lm(formula = `Amount Spent` ~ MarriedY, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
-904.1 -444.8 -145.5  283.2 3231.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   662.99      25.68   25.82  < 2e-16 ***
MarriedY      279.11      39.42    7.08 2.73e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 611 on 988 degrees of freedom
Multiple R-squared:  0.04829,	Adjusted R-squared:  0.04733
F-statistic: 50.13 on 1 and 988 DF,  p-value: 2.725e-12
```

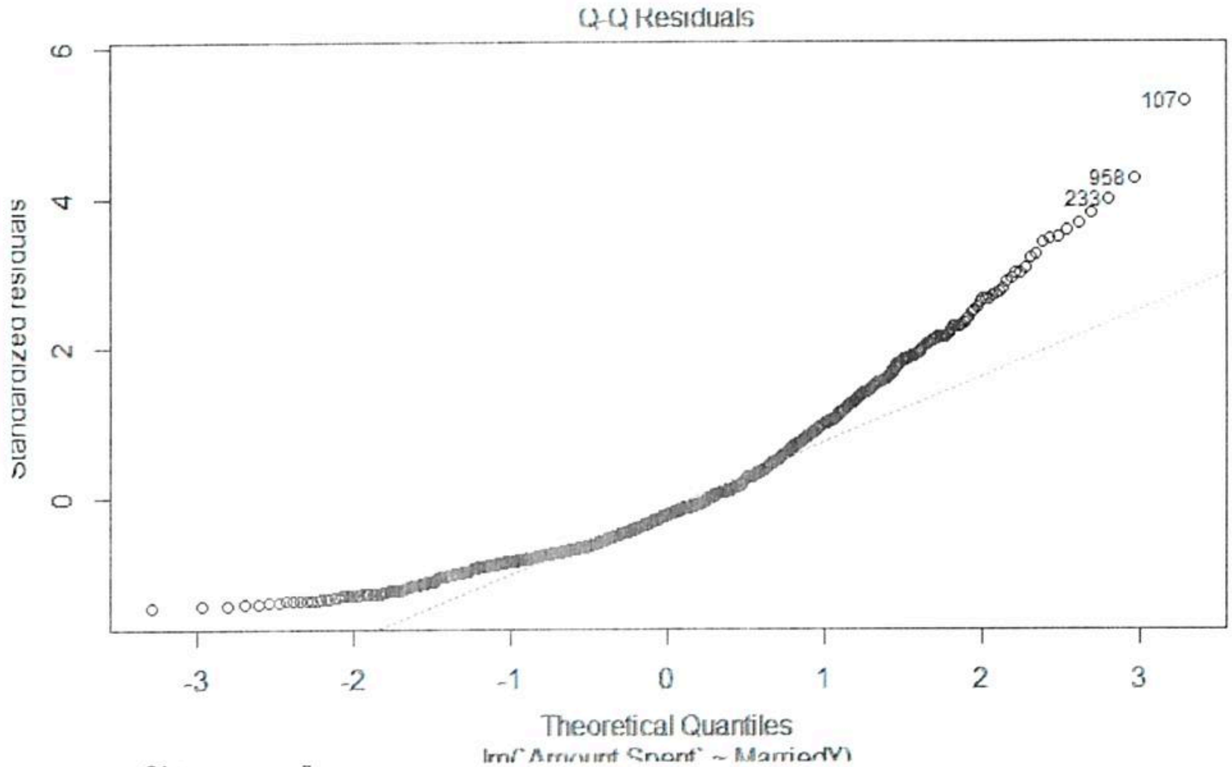
Machine found model:

```
Call:
lm(formula = `Amount Spent` ~ MarriedY + MediumInc + HighInc +
    HugeInc, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
-939.4 -429.7 -153.4  269.9 3261.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   616.15      34.44  17.890  < 2e-16 ***
MarriedY      164.25      59.29   2.770  0.00571 **
MediumInc      82.90      55.54   1.493  0.13584
HighInc       131.83      62.78   2.100  0.03600 *
HugeInc       209.01      78.84   2.651  0.00816 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 611.7 on 985 degrees of freedom
Multiple R-squared:  0.05535,	Adjusted R-squared:  0.05151
F-statistic: 14.43 on 4 and 985 DF,  p-value: 1.875e-11

Subset selection object
Call: regsubsets.formula(`Amount Spent` ~ ., data = data1, nvmax = 8,
    method = "seqrep")
8 Variables  (and intercept)
```

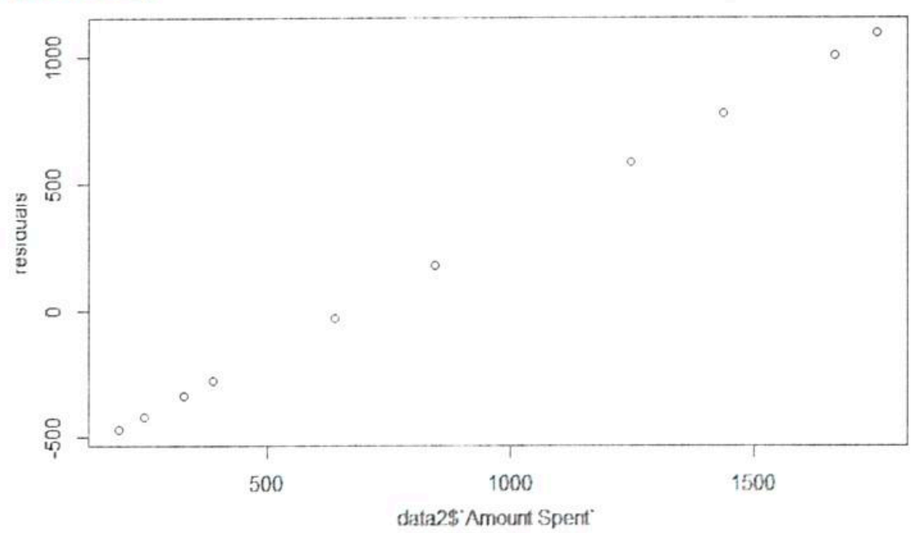
```
          Forced in Forced out
Middle        FALSE      FALSE
Older         FALSE      FALSE
Male          FALSE      FALSE
Own           FALSE      FALSE
MarriedY      FALSE      FALSE
MediumInc     FALSE      FALSE
HighInc       FALSE      FALSE
HugeInc       FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: 'sequential replacement'
         Middle Older Male Own MarriedY MediumInc HighInc HugeInc
1  ( 1 ) " "    " "   " "  " " "*"      " "       " "     " "
2  ( 1 ) "*"    "*"   " "  " " " "      " "       " "     " "
3  ( 1 ) " "    " "   " "  " " "*"      " "       "*"     "*"
4  ( 1 ) " "    " "   " "  " " "*"      "*"       "*"     "*"
5  ( 1 ) " "    " "   "*"  " " "*"      "*"       "*"     "*"
6  ( 1 ) " "    "*"   "*"  " " "*"      "*"       "*"     "*"
7  ( 1 ) " "    "*"   "*"  "*" "*"      "*"       "*"     "*"
8  ( 1 ) "*"    "*"   "*"  "*" "*"      "*"       "*"     "*"
```

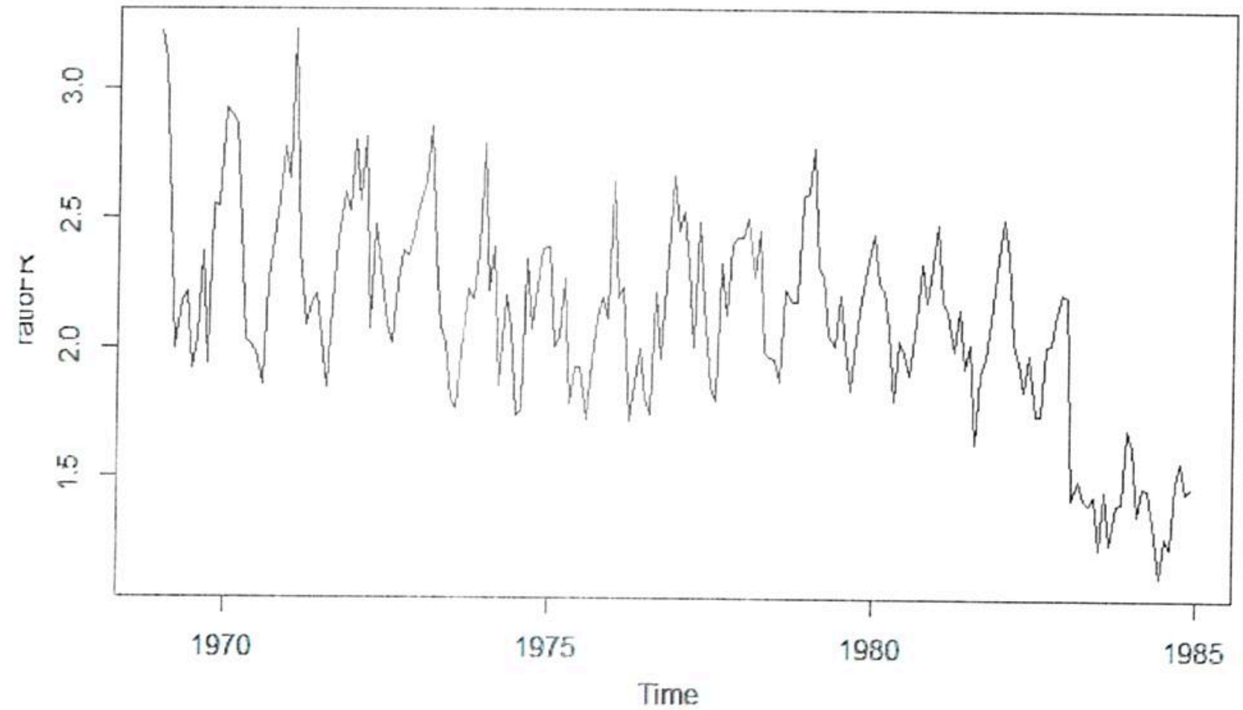


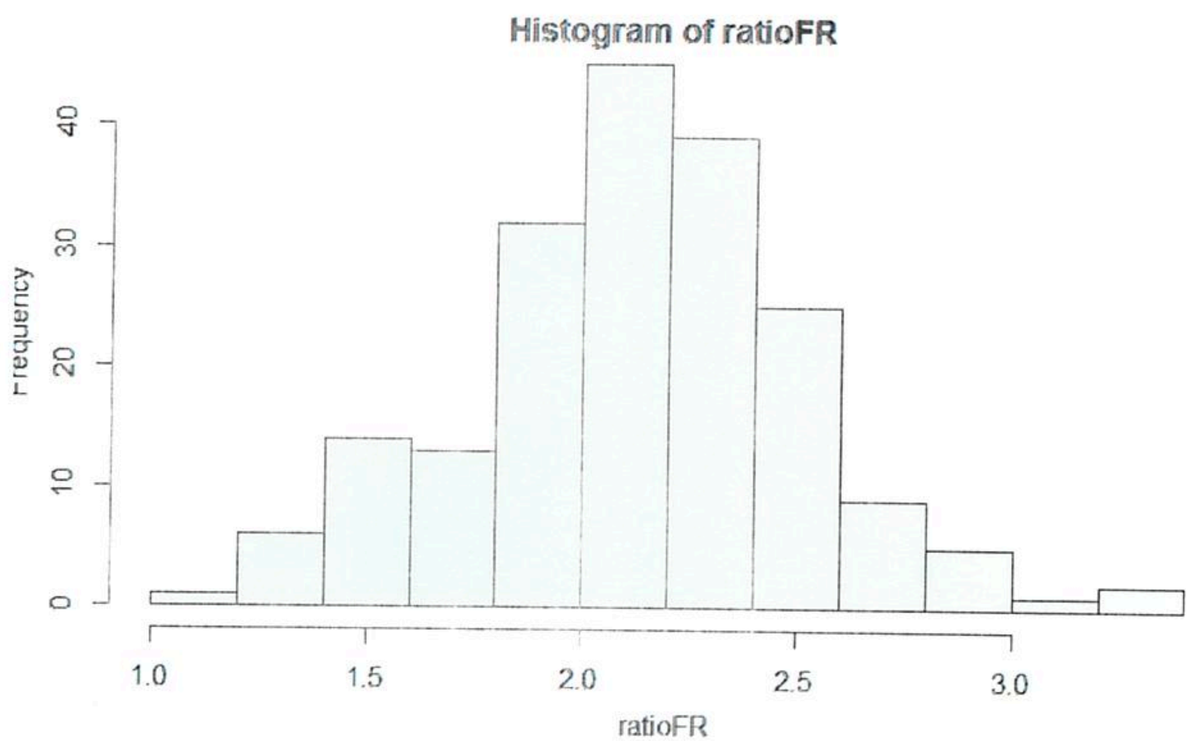


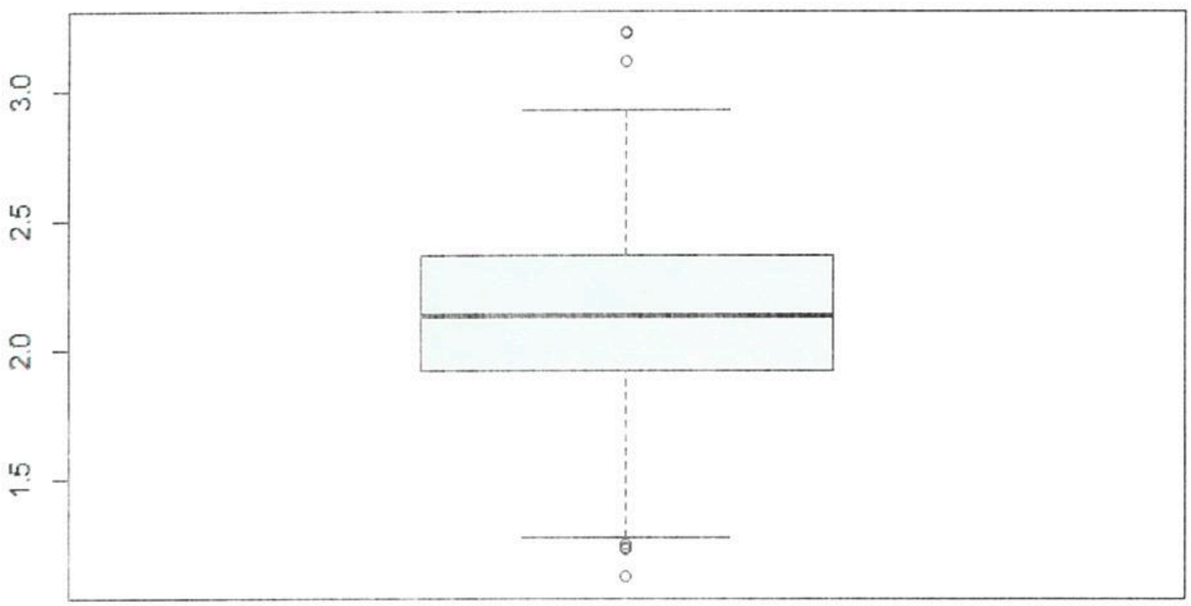


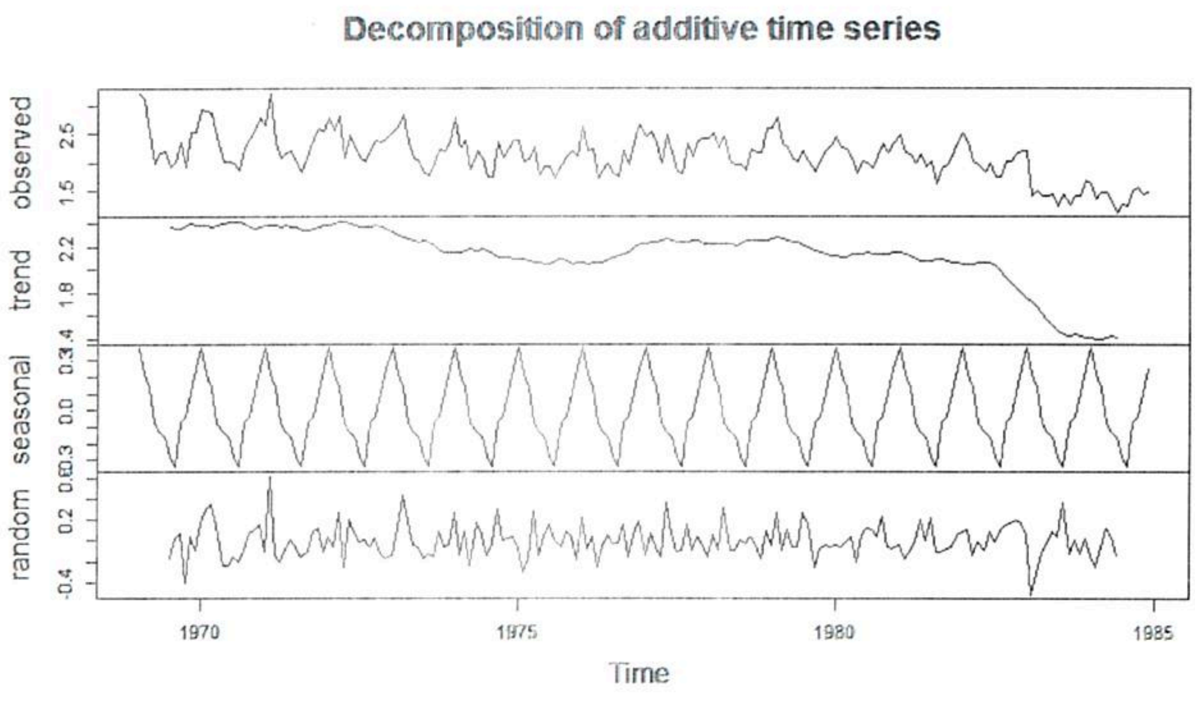

```
       fit       lwr      upr
1  662.9895 -541.0012 1866.980
2  662.9895 -541.0012 1866.980
3  662.9895 -541.0012 1866.980
4  662.9895 -541.0012 1866.980
5  942.1000 -262.2672 2146.467
6  662.9895 -541.0012 1866.980
7  942.1000 -262.2672 2146.467
8  942.1000 -262.2672 2146.467
9  662.9895 -541.0012 1866.980
10 942.1000 -262.2672 2146.467

246  842 1248  388  328  636  194 1668 1754 1438

612.9984
```

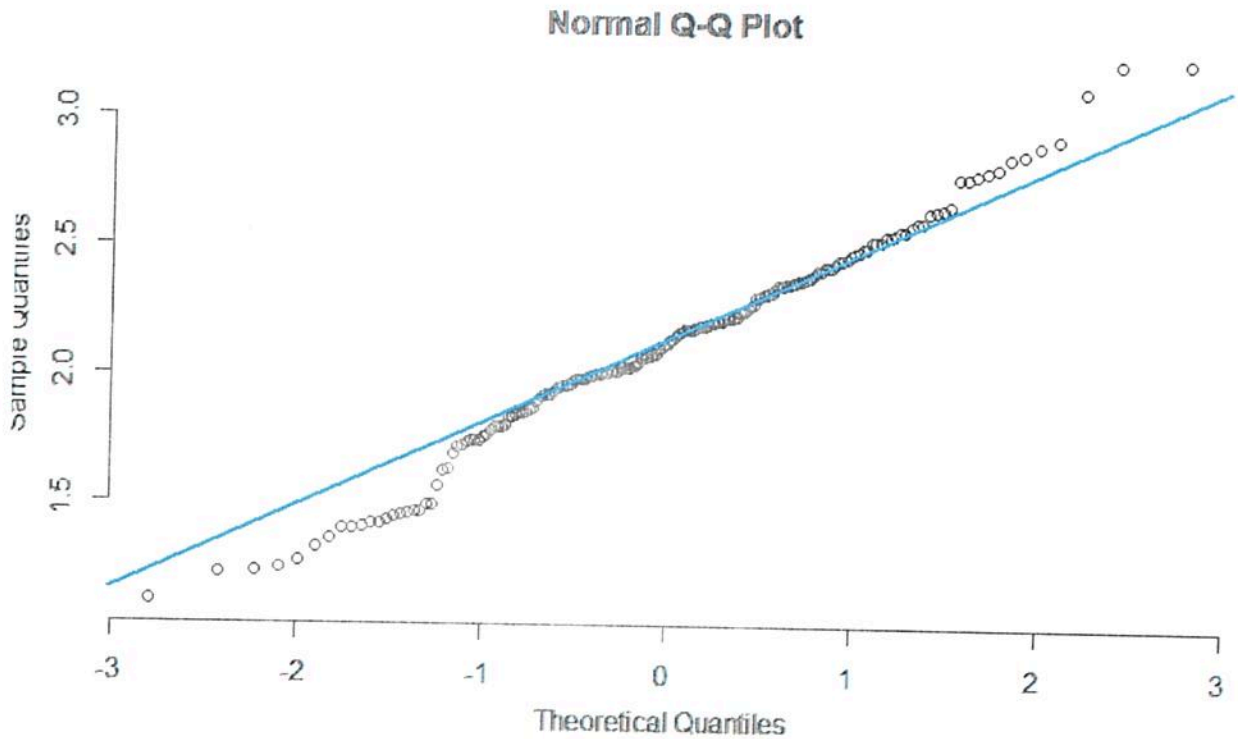



ratioFR
1.5
2.0
2.5
3.0
1970
1975
1980
1985
Time

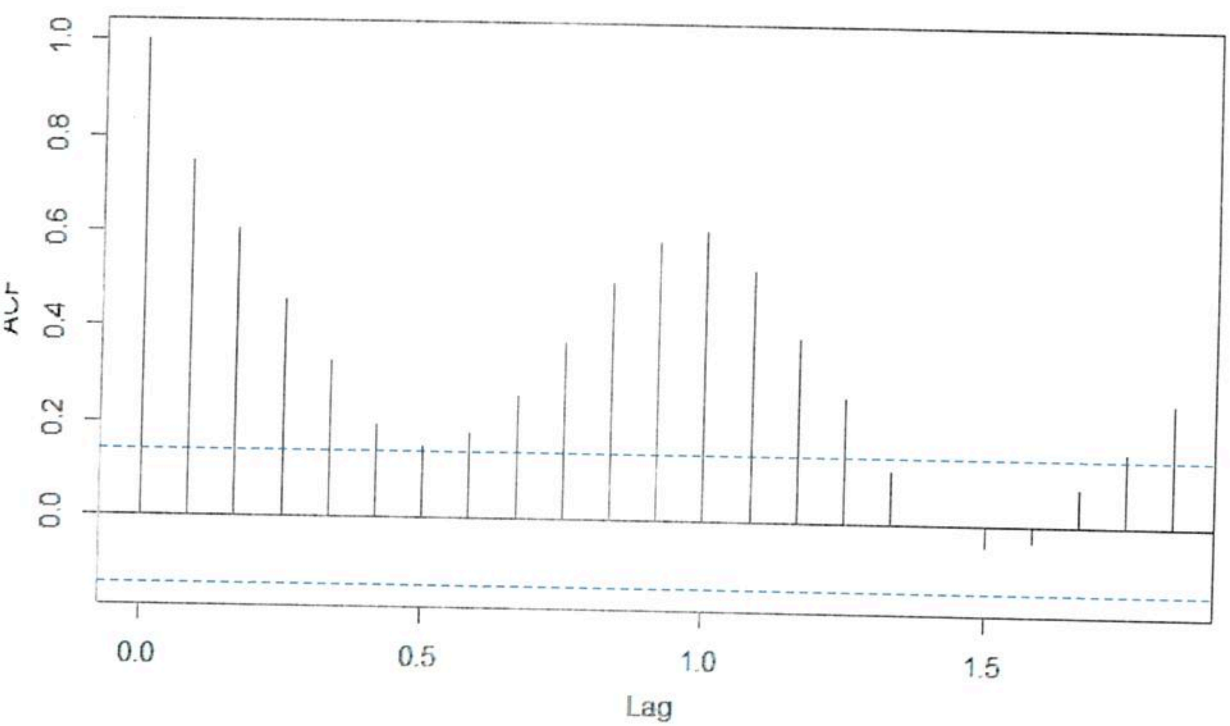

Histogram of ratioFR
Frequency
0
10
20
30
40
1.0
1.5
2.0
2.5
3.0
ratioFR


3.0
2.5
2.0
1.5

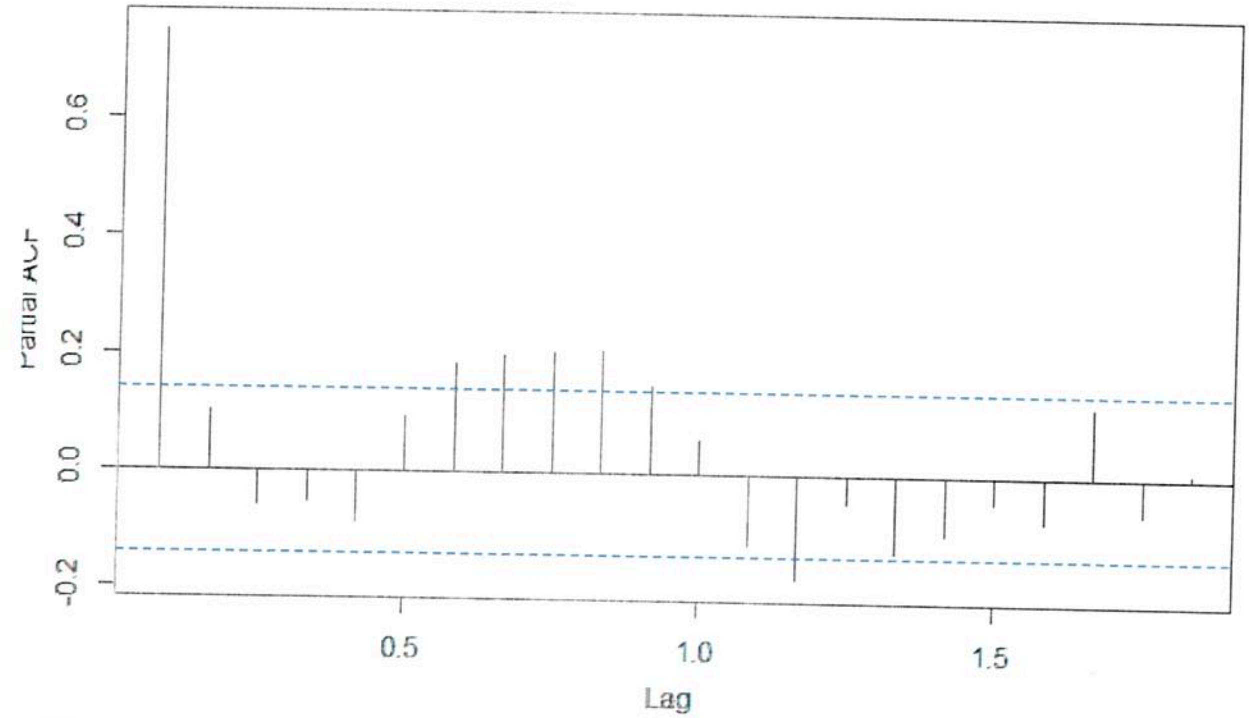

Decomposition of additive time series
observed
trend
seasonal
random
1970
1975
1980
1985
Time



Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles



Lag




```
Call:
arima(x = ratioFR, order = c(1, 1, 5))

Coefficients:
         ar1      ma1    ma2      ma3      ma4      ma5
      0.2623  -0.6308  0.089  -0.0650  -0.0689  -0.1855
s.e.  0.3138   0.3121  0.152   0.0901   0.1010   0.0843

sigma^2 estimated as 0.05585:  log likelihood = 4.05,  aic = 5.9
```