DSA301FinalExam_April2024

Benjamin

2024-05-29

SECTION A: IMPLEMENTATION QUESTIONS

Question A1 (Warmup) (1 point)

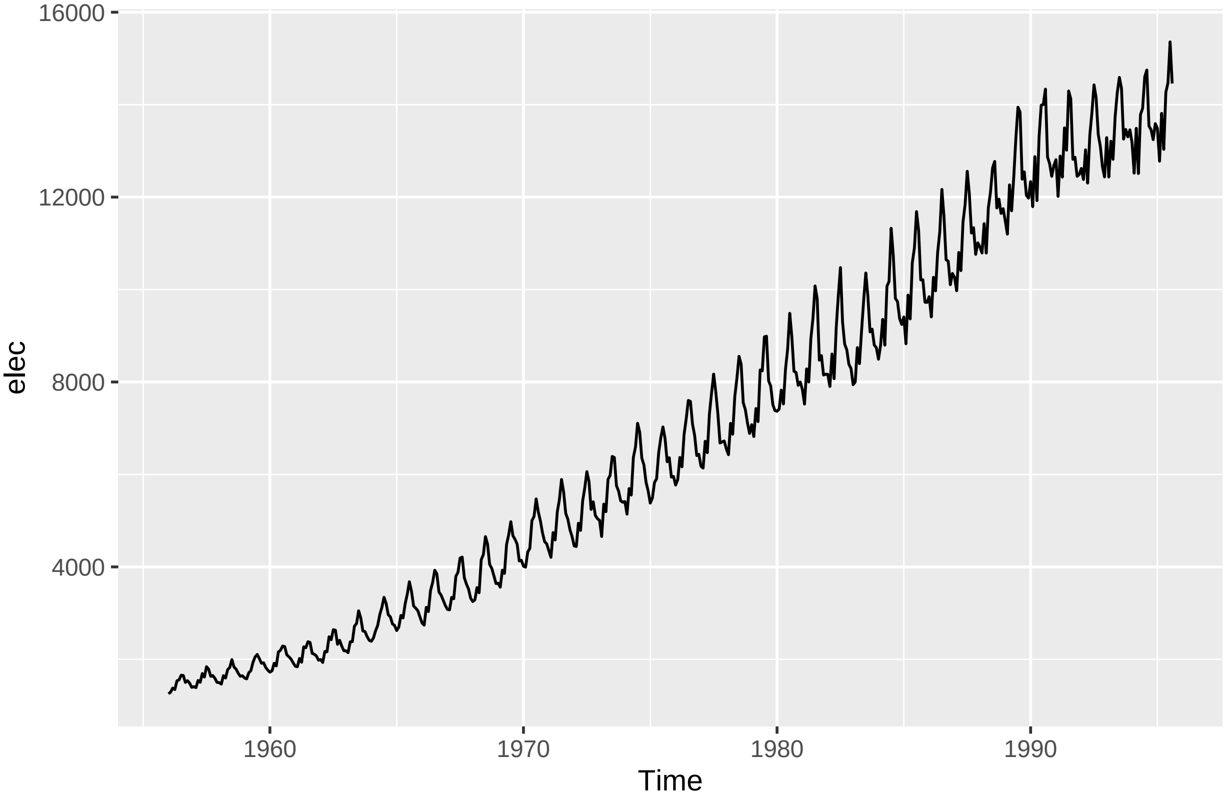
Fill in the code chunk for QUESTION 1 below to plot the classical time series decomposition of variable "elec" in fpp2. The line "autoplot(elec) is inserted as an example of how to generate plots in this environment. Feel free to delete it and/or replace with the correct answer.

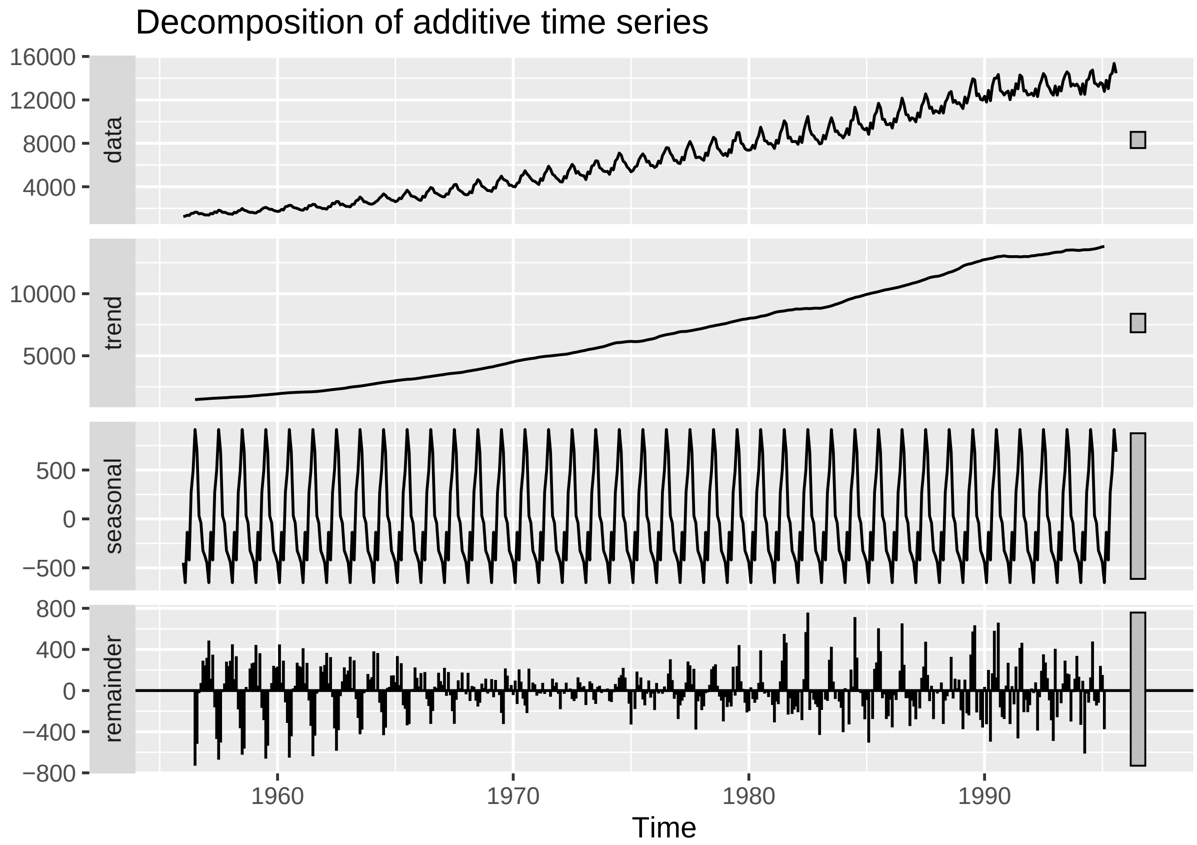
```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## -- Attaching packages ---------------------------------------------- fpp2 2.5 --
```

```
## v ggplot2   3.4.4      v fma       2.5
## v forecast  8.21.1     v expsmooth 2.3
```

```
##
```


16000
12000
8000
4000
elec
1960
1970
1980
1990
Time



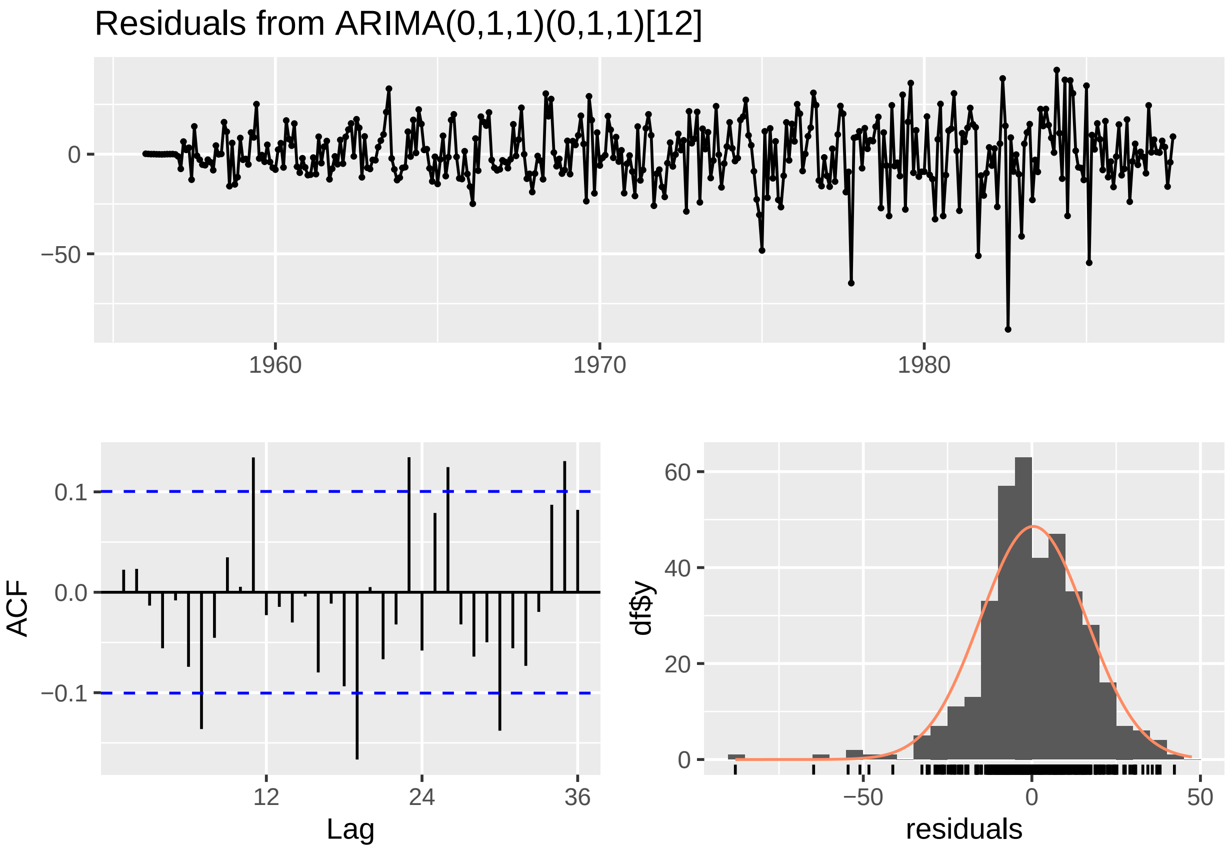
Question A2 (Benchmark Models) (2 points)

Fill in the code chunk for QUESTION 2 below to forecast the seasonal component of “elec” with the snaive method for 20 periods into the future. Feel free to delete/ replace any lines of code in the provided template which are not needed

```
##          Point Forecast      Lo 80      Hi 80      Lo 95      Hi 95
## Sep 1995       32.03529   32.03529   32.03529   32.03529   32.03529
## Oct 1995      -42.09184  -42.09184  -42.09184  -42.09184  -42.09184
## Nov 1995     -324.41129 -324.41129 -324.41129 -324.41129 -324.41129
## Dec 1995     -379.32902 -379.32902 -379.32902 -379.32902 -379.32902
## Jan 1996     -448.79697 -448.79697 -448.79697 -448.79697 -448.79697
## Feb 1996     -651.55552 -651.55552 -651.55552 -651.55552 -651.55552
## Mar 1996     -136.18659 -136.18659 -136.18659 -136.18659 -136.18659
## Apr 1996     -420.27321 -420.27321 -420.27321 -420.27321 -420.27321
## May 1996      274.00310  274.00310  274.00310  274.00310  274.00310
## Jun 1996      496.26516  496.26516  496.26516  496.26516  496.26516
## Jul 1996      913.85367  913.85367  913.85367  913.85367  913.85367
## Aug 1996      686.48722  686.48722  686.48722  686.48722  686.48722
## Sep 1996       32.03529   32.03529   32.03529   32.03529   32.03529
## Oct 1996      -42.09184  -42.09184  -42.09184  -42.09184  -42.09184
## Nov 1996     -324.41129 -324.41129 -324.41129 -324.41129 -324.41129
## Dec 1996     -379.32902 -379.32902 -379.32902 -379.32902 -379.32902
## Jan 1997     -448.79697 -448.79697 -448.79697 -448.79697 -448.79697
## Feb 1997     -651.55552 -651.55552 -651.55552 -651.55552 -651.55552
## Mar 1997     -136.18659 -136.18659 -136.18659 -136.18659 -136.18659
## Apr 1997     -420.27321 -420.27321 -420.27321 -420.27321 -420.27321
```

Question A3 (Model evaluation) (5 points)

Fill in the code chunk for QUESTION 3 below to automatically build an arima model on the seasonally adjusted component of elec. You may use built-in functions for automated order detection here. Leave a reasonable out of sample period, and compute accuracy statistics (e.g. RMSE, MAPE) of your model on the out of sample period. You do not need to preprocess the data for this question. Feel free to delete/ replace any lines of code in the provided template which are not needed


Residuals from ARIMA(0,1,1)(0,1,1)[12]

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)(0,1,1)[12]
## Q* = 48.424, df = 22, p-value = 0.0009535
##
## Model df: 2.   Total lags used: 24
```

```
##                       ME     RMSE       MAE         MPE     MAPE      MASE
## Training set    3.032536 130.9716  93.41836  0.09177956 1.891889 0.2943278
## Test set     -310.021346 746.9438 591.66705 -2.38991870 4.503117 1.8641309
##                      ACF1 Theil's U
## Training set -0.005775115        NA
## Test set      0.770809759  1.068048
```

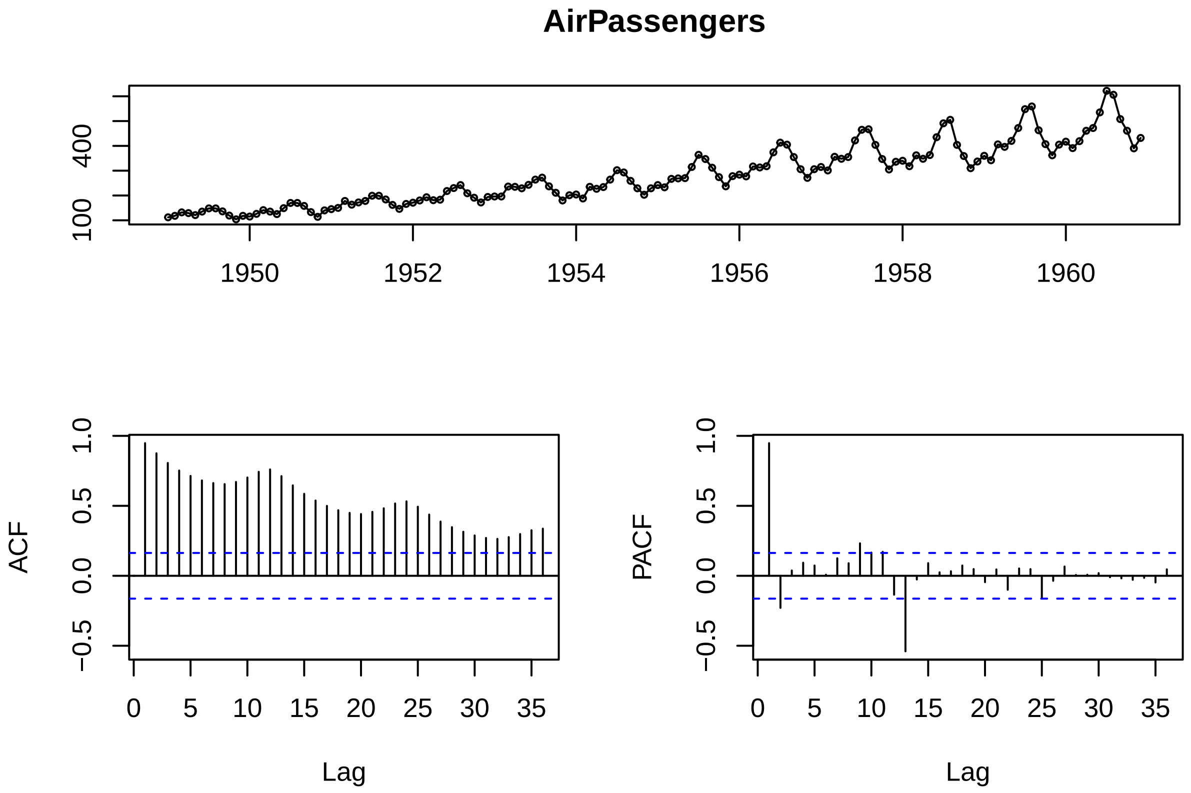
Question A4 (Preprocessing) (2 point)

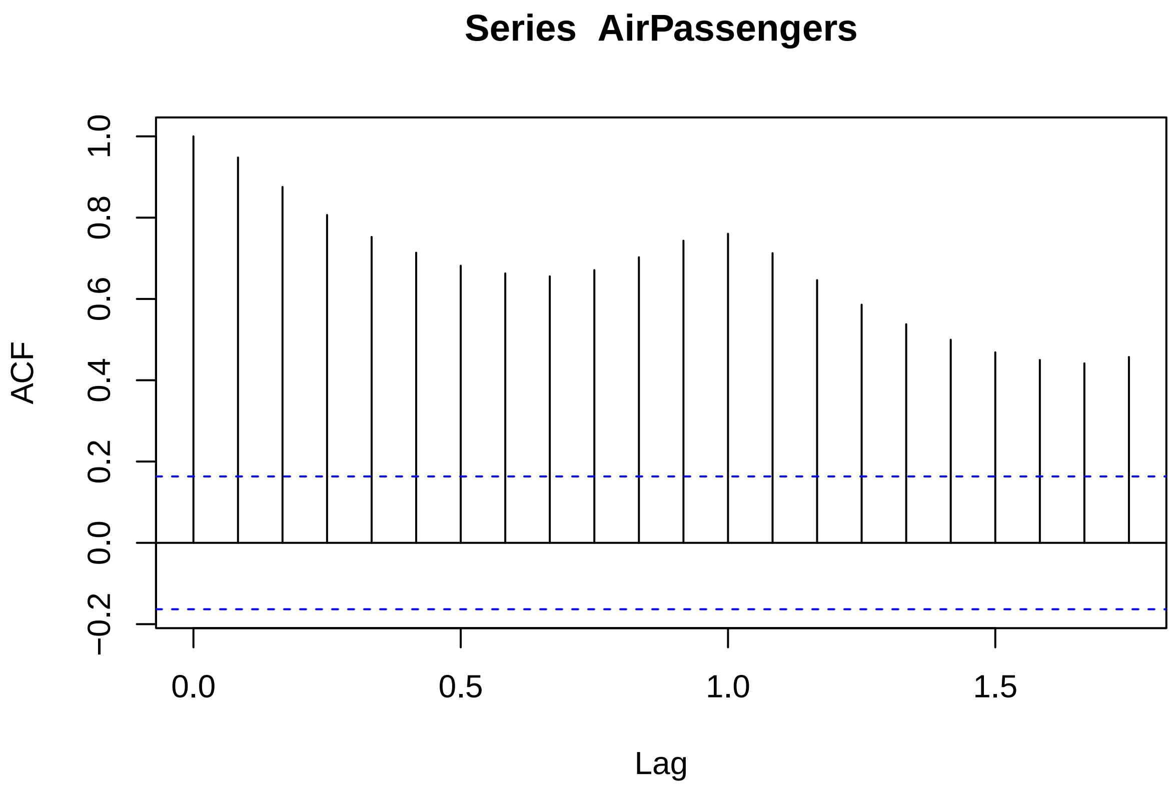
Fill in the code chunk for QUESTION 4 below to correct for variance in elec due to mechanical differences in the number of days each month. Delete sample code lines which are not required.

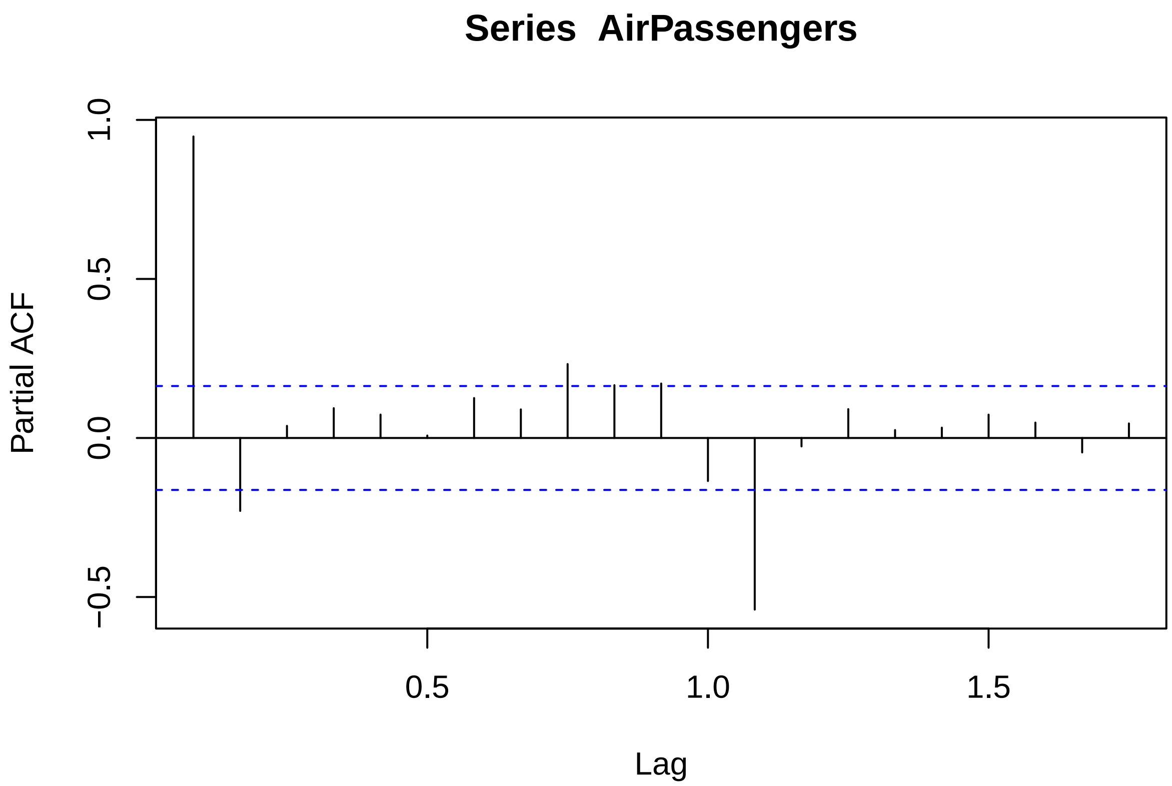
```
##            Jan       Feb       Mar       Apr       May       Jun       Jul
## 1956  40.45161  44.48276  44.48387  44.86667  49.51613  51.83333  53.38710
## 1957  45.45161  49.53571  49.77419  50.06667  54.61290  53.86667  59.38710
## 1958  48.29032  52.25000  53.16129  53.16667  57.32258  60.80000  64.32258
## 1959  51.51613  56.32143  55.12903  58.53333  62.45161  68.40000  67.90323
## 1960  55.51613  60.41379  61.74194  61.90000  69.64516  73.16667  73.77419
## 1961  59.70968  65.67857  65.12903  64.56667  73.22581  75.03333  76.83871
## 1962  64.35484  69.00000  70.03226  72.06667  80.29032  80.80000  85.19355
## 1963  70.45161  76.57143  76.74194  79.43333  87.64516  92.46667  98.41935
## 1964  77.09677  84.93103  84.38710  91.13333  95.80645 104.16667 107.80645
## 1965  84.58065  96.35714  95.16129  96.50000 103.22581 113.60000 118.67742
## 1966  89.87097  97.82143 100.80645 101.10000 112.45161 122.03333 126.67742
## 1967  99.35484 109.60714 107.74194 110.33333 122.51613 129.43333 135.19355
## 1968 104.83871 113.34483 114.58065 114.66667 133.96774 142.16667 150.16129
## 1969 117.64516 127.14286 126.74194 128.60000 144.67742 156.56667 160.54839
## 1970 129.48387 142.64286 139.35484 146.66667 161.35484 169.70000 176.48387
## 1971 140.32258 150.21429 153.00000 152.73333 167.45161 181.90000 190.03226
## 1972 143.64516 153.10345 159.51613 159.60000 175.00000 190.20000 195.51613
## 1973 161.54839 166.32143 172.87097 173.10000 190.03226 199.33333 206.12903
## 1974 174.61290 183.60714 183.70968 185.13333 205.45161 219.73333 229.25806
## 1975 173.51613 196.03571 187.87097 196.90000 209.09677 226.50000 226.70968
## 1976 186.09677 203.00000 205.38710 205.50000 221.54839 240.03333 245.19355
## 1977 199.22581 219.21429 216.67742 215.66667 235.87097 258.76667 263.58065
## 1978 211.35484 229.53571 229.19355 228.96667 247.83871 269.40000 275.96774
## 1979 228.29032 243.57143 239.54839 238.10000 266.48387 274.66667 289.58065
## 1980 237.61290 255.65517 252.38710 250.80000 267.06452 290.23333 306.00000
## 1981 252.70968 268.60714 267.22581 266.63333 288.38710 312.70000 325.09677
## 1982 263.41935 282.25000 277.61290 269.03333 296.06452 329.10000 337.93548
## 1983 256.19355 285.75000 282.06452 279.90000 294.03226 325.76667 334.12903
## 1984 273.93548 303.27586 301.74194 293.20000 324.90323 339.13333 365.35484
## 1985 303.45161 315.25000 318.70968 312.13333 341.29032 363.30000 377.00000
## 1986 317.61290 335.96429 331.12903 332.33333 348.41935 374.86667 392.48387
## 1987 331.06452 356.17857 348.48387 346.96667 369.61290 394.83333 405.12903
## 1988 352.35484 372.06897 368.61290 359.60000 379.74194 403.46667 407.54839
## 1989 370.48387 399.92857 395.64516 390.13333 400.61290 441.96667 449.83871
## 1990 397.93548 421.17857 415.38710 397.43333 429.22581 466.26667 451.67742
## 1991 413.22581 429.10714 415.74194 414.36667 435.45161 433.80000 461.16129
## 1992 407.12903 426.89655 420.09677 410.06667 430.29032 460.83333 465.41935
## 1993 428.61290 444.07143 426.09677 427.23333 443.41935 475.30000 470.64516
## 1994 424.87097 447.03571 435.12903 416.96667 444.67742 464.03333 471.06452
## 1995 435.06452 456.28571 445.54839 434.40000 460.25806 482.43333 495.45161
##            Aug       Sep       Oct       Nov       Dec
## 1956  53.25806  50.00000  49.61290  49.53333  44.96774
## 1957  57.64516  54.36667  53.19355  52.86667  48.38710
## 1958  59.19355  59.56667  54.80645  54.43333  53.06452
## 1959  65.03226  63.80000  62.09677  60.80000  56.93548
## 1960  73.41935  69.86667  66.29032  66.80000  62.06452
## 1961  76.25806  70.96667  68.06452  69.06667  63.87097
## 1962  84.83871  77.46667  77.80645  76.13333  70.51613
## 1963  93.25806  87.10000  83.87097  83.10000  77.74194
## 1964 103.45161  98.80000  94.16129  92.13333  88.12903
## 1965 112.03226 105.13333 100.22581 101.73333  94.12903
## 1966 124.22581 115.20000 109.35484 109.33333 102.12903
## 1967 135.90323 125.53333 117.03226 117.33333 107.16129
## 1968 144.90323 135.03333 127.96774 126.90000 117.38710
## 1969 150.80645 153.20000 144.87097 137.56667 133.67742
## 1970 167.51613 166.56667 152.80645 151.53333 145.09677
## 1971 181.22581 171.93333 162.25806 160.00000 150.12903
## 1972 188.58065 174.73333 174.45161 170.46667 162.64516
## 1973 205.35484 191.86667 181.93548 180.96667 174.12903
## 1974 223.12903 211.76667 200.16129 194.33333 182.12903
## 1975 218.58065 209.13333 205.22581 198.00000 192.19355
## 1976 244.54839 236.33333 220.67742 213.60000 207.58065
## 1977 251.22581 243.70000 215.45161 223.46667 216.90323
## 1978 270.51613 251.76667 238.64516 237.06667 222.12903
## 1979 290.03226 267.53333 255.19355 250.33333 238.09677
## 1980 289.45161 274.36667 264.70968 264.23333 258.03226
## 1981 316.00000 282.36667 276.51613 271.66667 263.48387
## 1982 299.87097 293.93333 280.54839 279.36667 267.51613
## 1983 317.70968 302.76667 294.93548 293.33333 281.96774
## 1984 346.58065 326.86667 314.19355 312.43333 298.19355
## 1985 363.87097 340.26667 329.41935 324.16667 313.58065
## 1986 373.48387 354.83333 342.35484 336.80000 333.80645
## 1987 389.35484 374.03333 365.74194 358.70000 355.22581
## 1988 412.00000 392.13333 385.67742 388.20000 379.03226
## 1989 446.41935 412.90000 404.70968 401.26667 386.35484
## 1990 462.45161 428.90000 410.35484 414.96667 409.22581
## 1991 455.64516 427.23333 414.90323 414.96667 402.87097
## 1992 456.48387 445.16667 422.38710 421.86667 401.12903
## 1993 463.03226 441.80000 434.32258 443.40000 434.06452
## 1994 475.77419 451.33333 434.09677 441.43333 438.38710
## 1995 466.35484
```

Question A5 (Arima modelling "by hand") (12 points)

Fill in the code chunk below to: a. Plot ACF and PACF of AirPassengers. For this question, disregard time series decomposition b. Form reasonable hypothesis on what ARIMA orders the variable should take. Remember to check ndiffs and nsdiffs c. Test the hypotheses so formed from above, and select the best one based on the appropriate model selection criteria. Keep an out of sample for this step d. Evaluate the performance of your model out of sample. e. What else do we need to do to ensure that the model has sufficiently exploited all time series information?


AirPassengers
100
400
1950
1952
1954
1956
1958
1960
ACF
−0.5
0.0
0.5
1.0
0
5
10
15
20
25
30
35
Lag
PACF
−0.5
0.0
0.5
1.0
0
5
10
15
20
25
30
35
Lag


Series AirPassengers
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.5
1.0
1.5
Lag

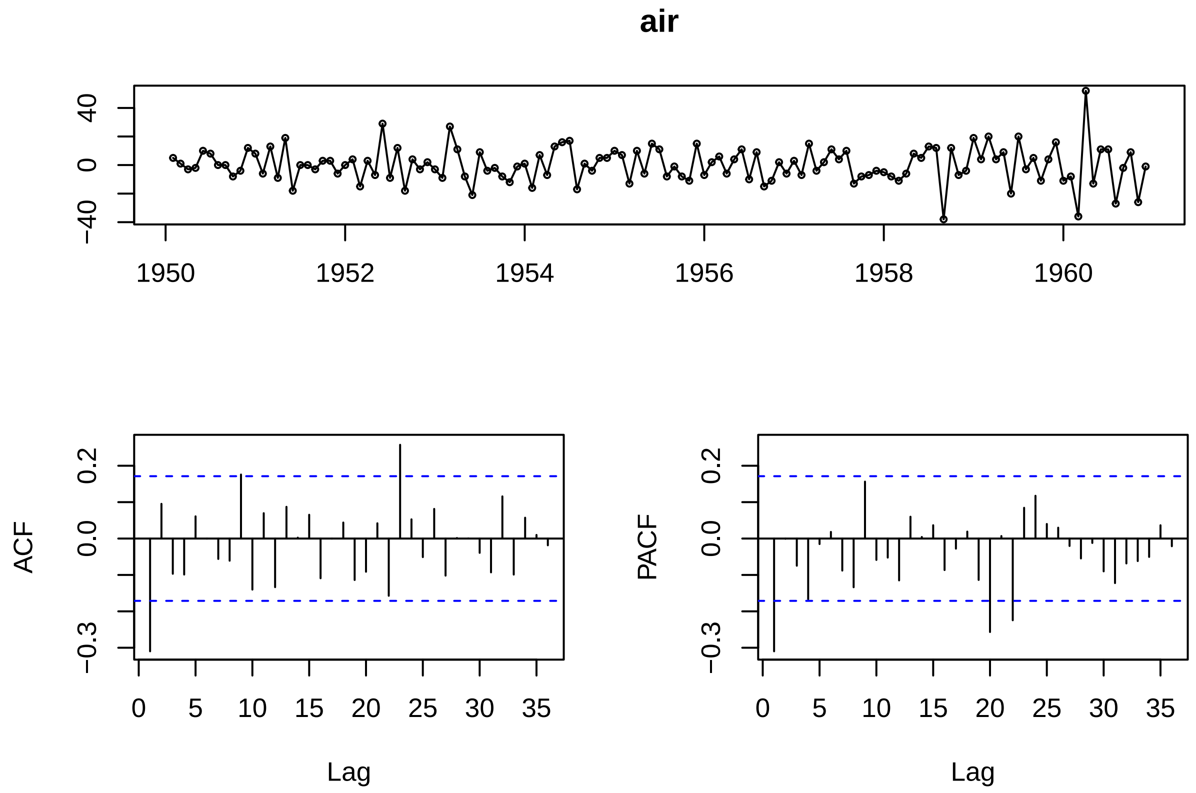


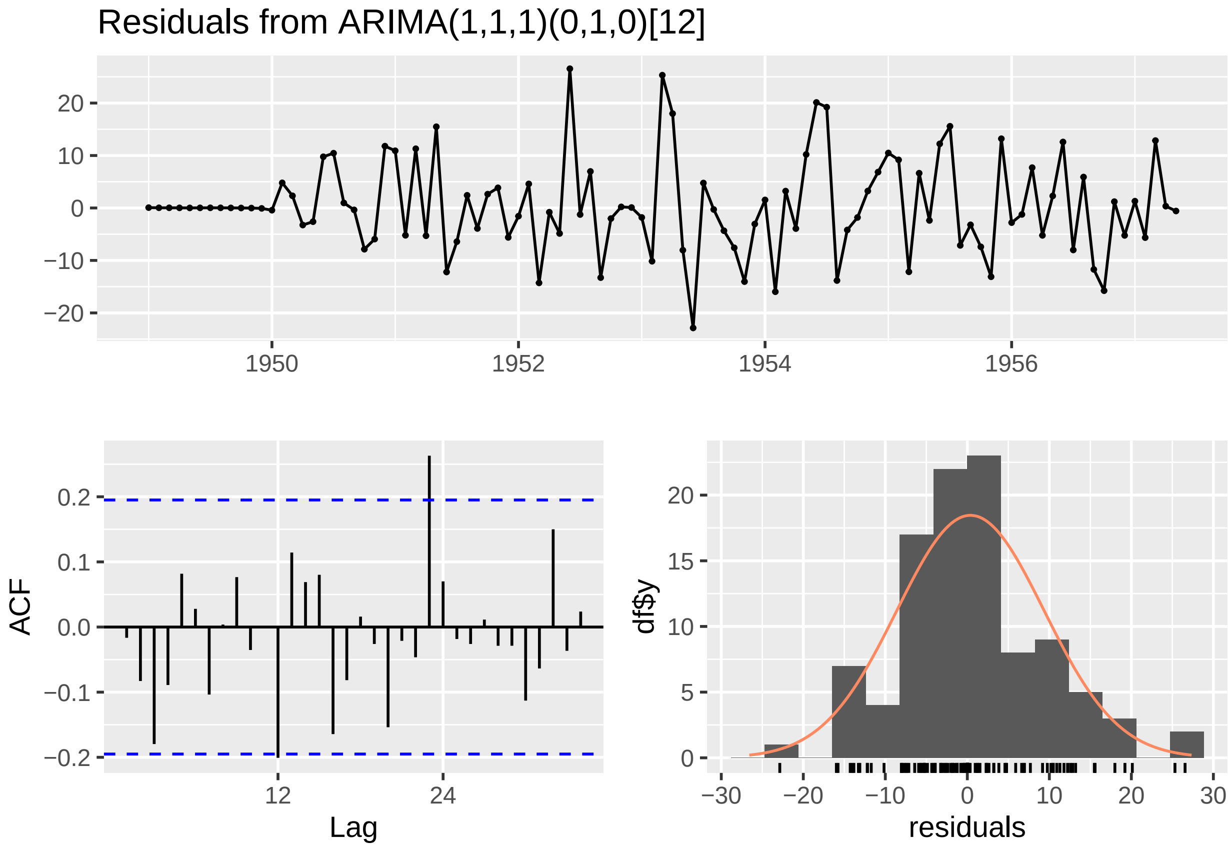
```
## [1] 1

## [1] 0

## [1] 1

##
## #######################
## # KPSS Unit Root Test #
## #######################
##
## Test is of type: mu with 4 lags.
##
## Value of test-statistic is: 0.0543
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

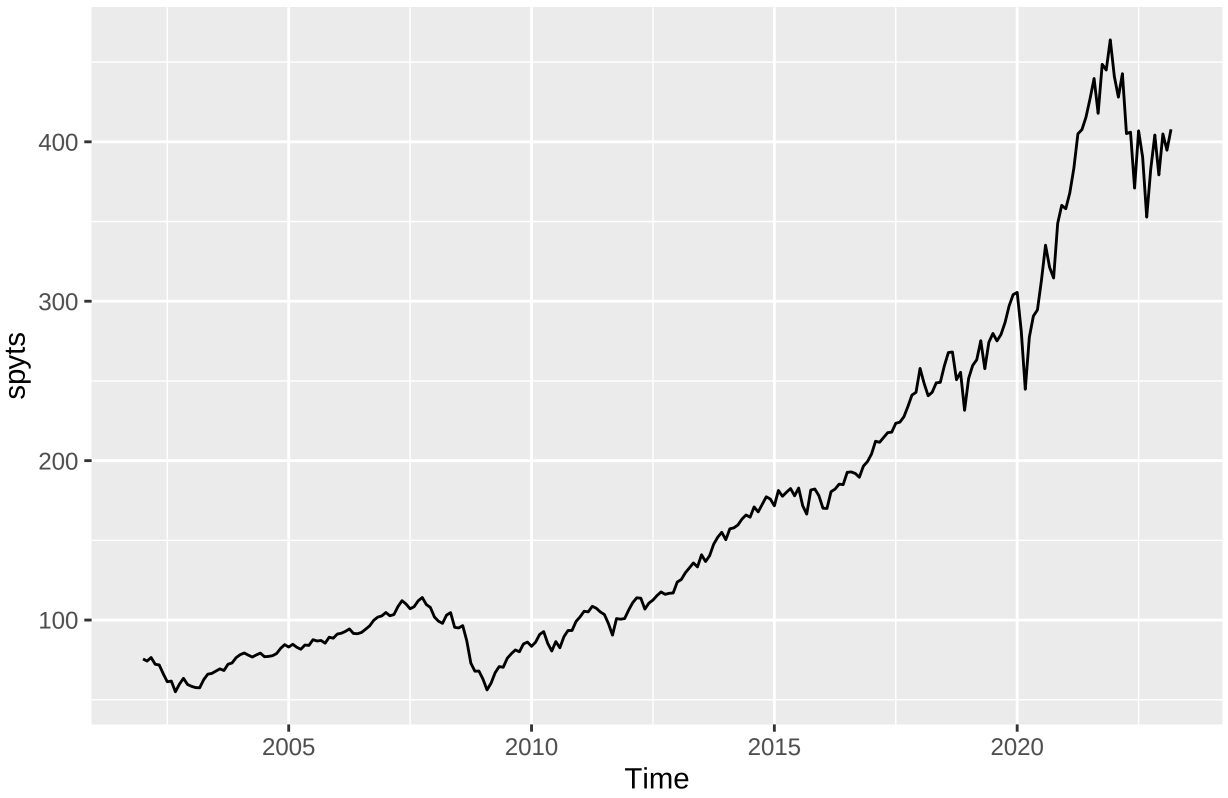

air
40
0
−40
1950
1952
1954
1956
1958
1960
ACF
0.2
0.0
−0.3
0
5
10
15
20
25
30
35
Lag
PACF
0.2
0.0
−0.3
0
5
10
15
20
25
30
35
Lag



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,1)(0,1,0)[12]
## Q* = 22.726, df = 18, p-value = 0.2013
##
## Model df: 2.   Total lags used: 20
```

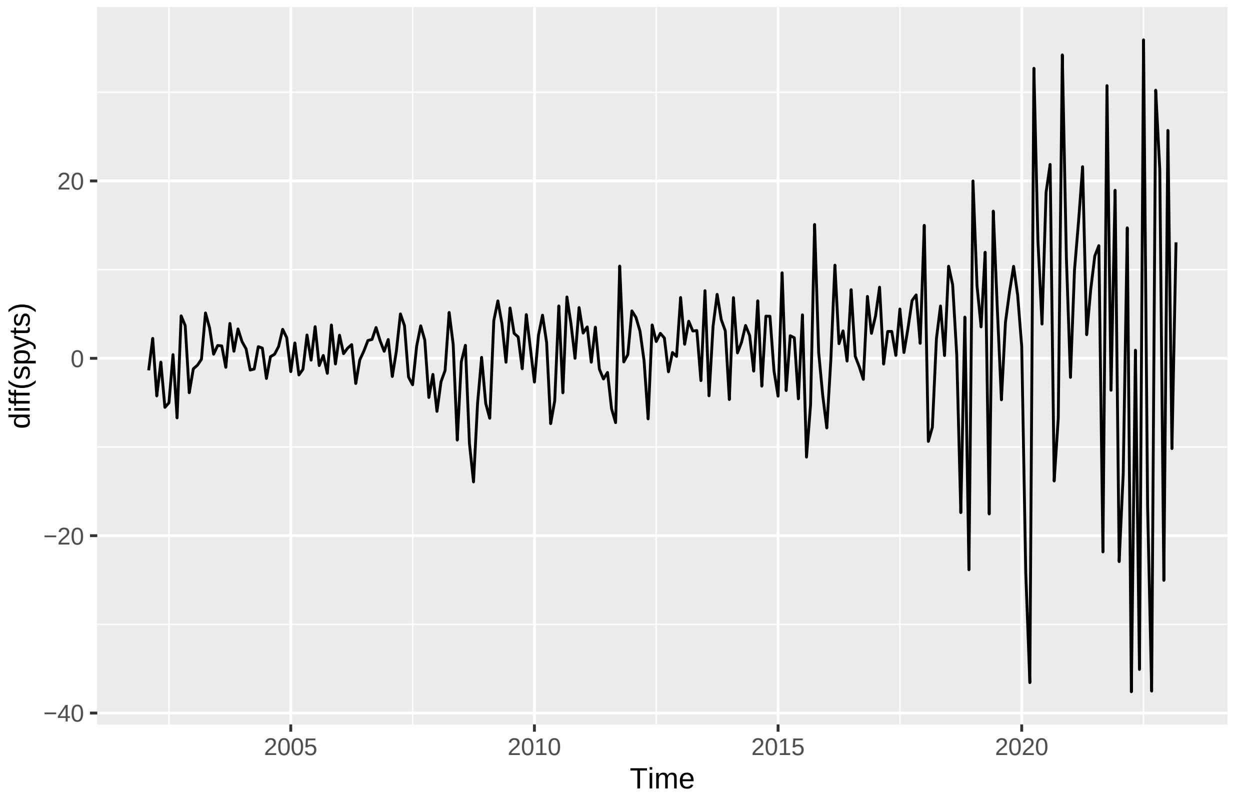
```
##                      ME      RMSE       MAE        MPE     MAPE      MASE
## Training set  0.3801447  8.951234  6.533904  0.1152067 2.950961 0.2220380
## Test set     -4.2384465 25.040239 20.430268 -1.8069438 4.870793 0.6942703
##                     ACF1 Theil's U
## Training set -0.01652296        NA
## Test set      0.63191004 0.5357249
```

##Question A6 (Free form modelling) (18 points) Extend the code chunk below to build the best possible model of SPY (S&P 500) prices that you can. You can edit the code to open data for other equities instruments if you like, from the set that was downloaded There is (deliberately) less structure to this question, please exercise independent judgement. There is also no single correct answer, although responses will be graded exclusively on how consistent the approach is to 'best practice' and how complete the exploration is. NOTE: make sure 'SPY.csv', which you downloaded from eLearn is in the same folder. Otherwise just put the fully qualified path to the file, with ("//") separating directories.



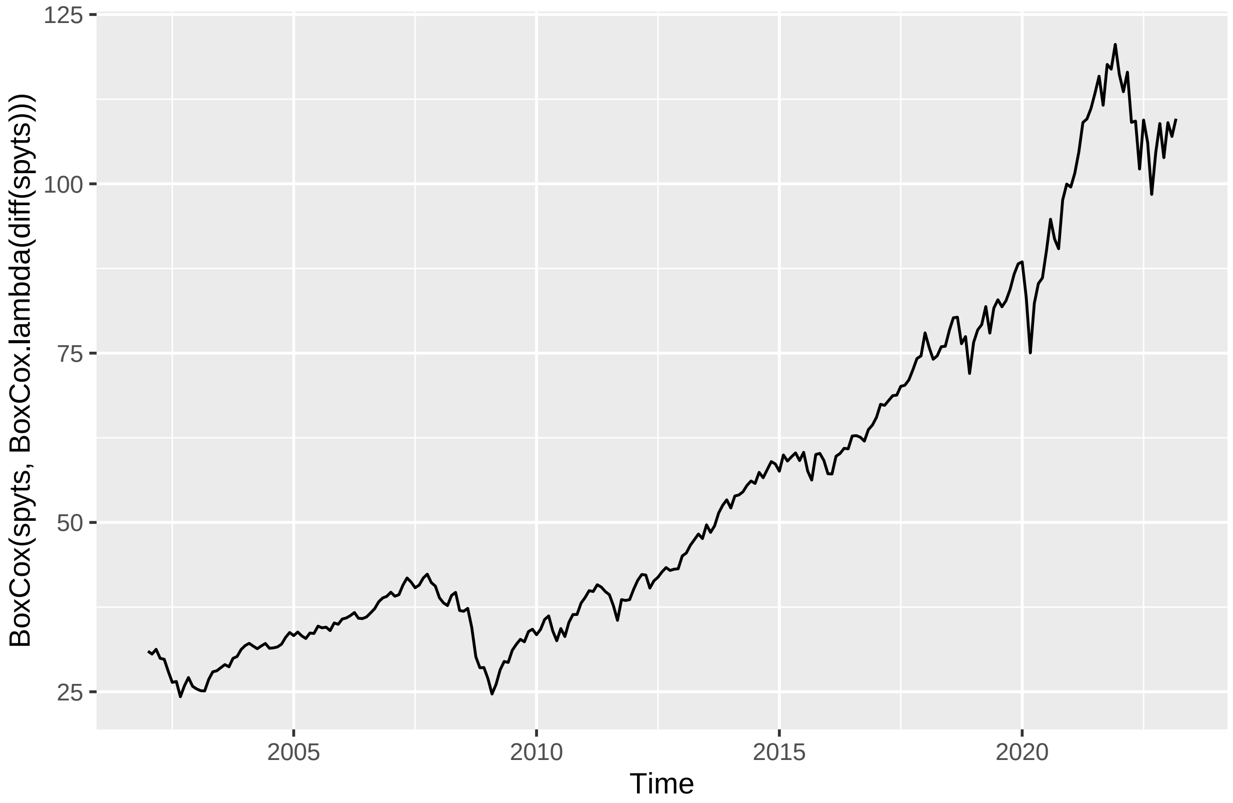
```
## [1] 0
```

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

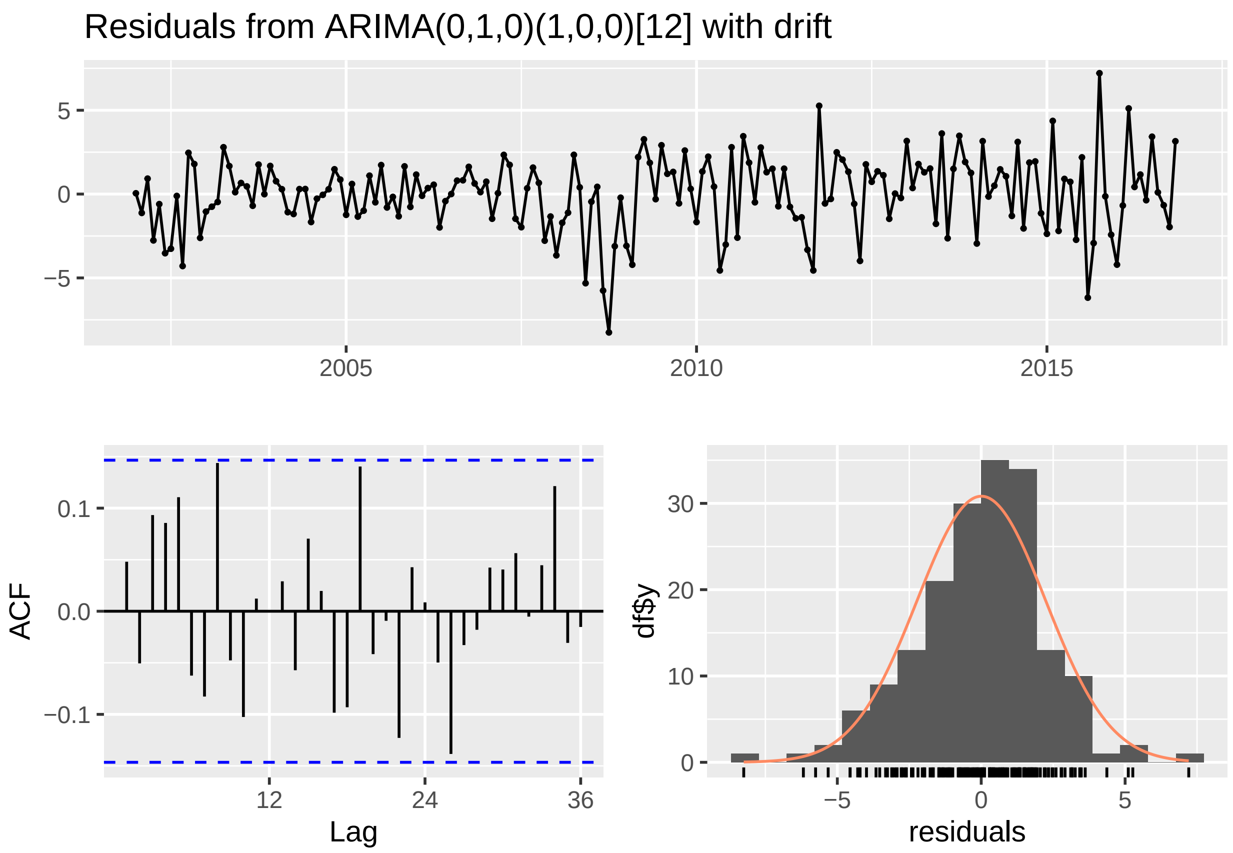


```
##
## #######################
## # KPSS Unit Root Test #
## #######################
##
## Test is of type: mu with 5 lags.
##
## Value of test-statistic is: 0.4005
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

```
## Warning in guerrero(x, lower, upper): Guerrero's method for selecting a Box-Cox
## parameter (lambda) is given for strictly positive data.
```



125
100
75
50
25
BoxCox(spyts, BoxCox.lambda(diff(spyts)))
2005
2010
2015
2020
Time

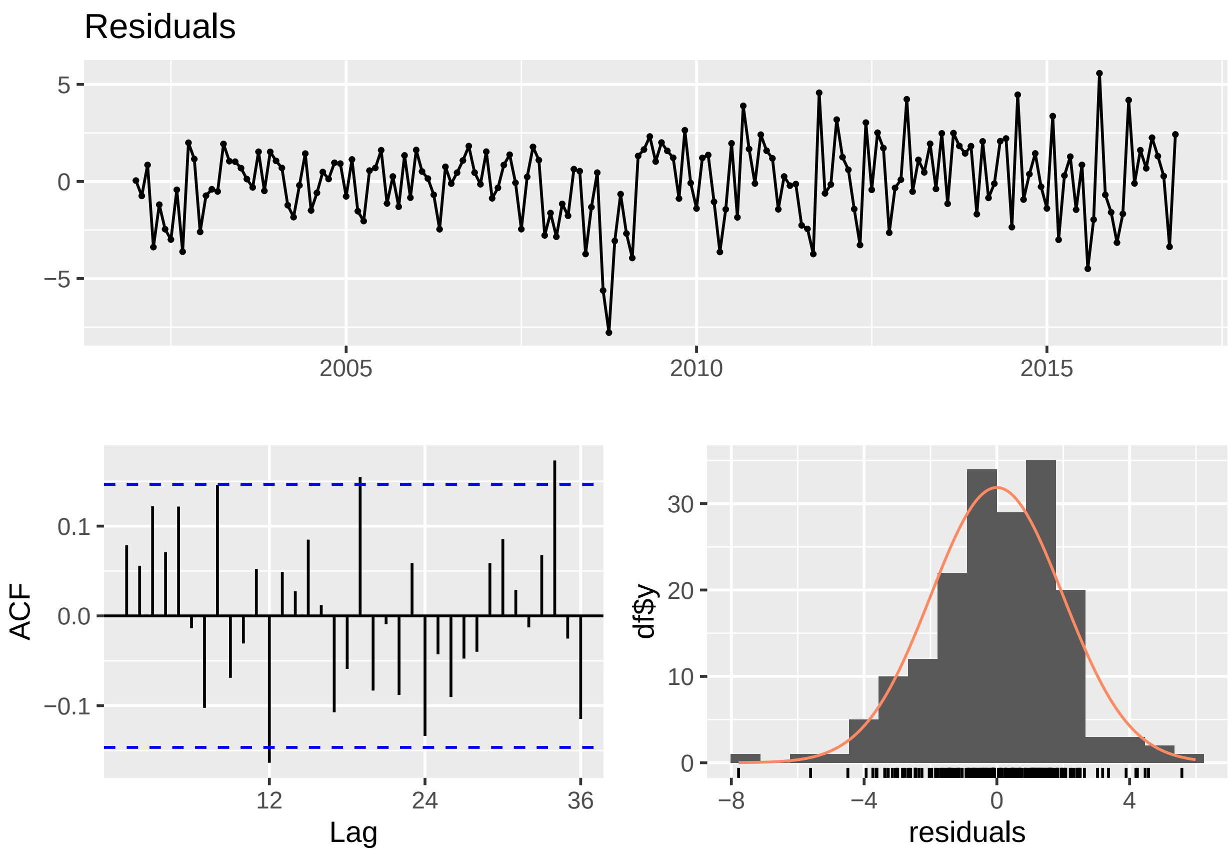


```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,0)(1,0,0)[12] with drift
## Q* = 27.946, df = 23, p-value = 0.2178
##
## Model df: 1.   Total lags used: 24
```

```
##                        ME       RMSE       MAE        MPE      MAPE      MASE
## Training set  0.02780304   4.133146  3.149882 -0.199665  3.133225 0.2087965
## Test set     89.75006586 109.517187 89.750066 25.429212 25.429212 5.9492713
##                    ACF1 Theil's U
## Training set 0.03038941        NA
## Test set     0.94493599  5.709334
```

```
##                        ME       RMSE        MAE         MPE      MAPE      MASE
## Training set   0.4311004   3.658946   2.874615  0.3528742  2.857801 0.1905499
## Test set     114.1636120 137.434311 114.163612 32.5155247 32.515525 7.5675744
##                     ACF1 Theil's U
## Training set -0.02810118        NA
## Test set      0.95384282  7.225452
```

```
##                        ME       RMSE       MAE        MPE      MAPE      MASE
## Training set  0.02075779   3.674009  2.868194 -0.2023003  2.87461 0.1901243
## Test set     91.70349252 111.209726 91.703493 26.0737168 26.07372 6.0787583
##                    ACF1 Theil's U
## Training set 0.05836171        NA
## Test set     0.94639239  5.815805
```



```
##
##  Ljung-Box test
##
## data:  Residuals
## Q* = 38.399, df = 24, p-value = 0.03154
##
## Model df: 0.   Total lags used: 24

## Warning: package 'stringr' was built under R version 4.3.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## Joining with `by = join_by(Date)`
## Joining with `by = join_by(Date)`

##
## ######################
## # Johansen-Procedure #
## ######################
##
## Test type: maximal eigenvalue statistic (lambda max) , with linear trend
```

```
##
## Eigenvalues (lambda):
## [1] 1.053015e-01 2.220913e-02 3.263406e-05
##
## Values of teststatistic and critical values of test:
##
##           test 10pct  5pct  1pct
## r <= 2 |  0.01  6.50  8.18 11.65
## r <= 1 |  3.98 12.91 14.90 19.19
## r = 0  | 19.69 18.90 21.07 25.75
##
## Eigenvectors, normalised to first column:
## (These are the cointegration relations)
##
##                  Adj.Close.l2 xlf_adj_close.l2 qqq_adj_close.l2
## Adj.Close.l2         1.000000         1.000000         1.000000
## xlf_adj_close.l2    -1.496335         4.804744       -10.442522
## qqq_adj_close.l2    -1.380368        -2.317076        -2.758383
##
## Weights W:
## (This is the loading matrix)
##
##                 Adj.Close.l2 xlf_adj_close.l2 qqq_adj_close.l2
## Adj.Close.d       0.02184345     -0.017935436    -2.106056e-04
## xlf_adj_close.d  -0.02855972     -0.003251011    -1.006811e-05
## qqq_adj_close.d   0.09060609     -0.011499363    -5.471430e-05
```

```
## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## The following objects are masked from 'package:fma':
##
##     cement, housing, petrol
##
## Loading required package: strucchange
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Loading required package: sandwich
##
## Attaching package: 'strucchange'
##
## The following object is masked from 'package:stringr':
##
```

```
##     boundary
##
## Loading required package: lmtest
```

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      1      1      1      1
##
## $criteria
##              1        2        3        4        5        6        7        8
## AIC(n) 1.339820 1.351532 1.348415 1.370793 1.346824 1.403369 1.475705 1.494937
## HQ(n)  1.430010 1.509364 1.573889 1.663910 1.707583 1.831770 1.971749 2.058623
## SC(n)  1.562061 1.740454 1.904018 2.093077 2.235789 2.459016 2.698032 2.883945
## FPE(n) 3.818457 3.863889 3.852914 3.942073 3.851760 4.080555 4.393660 4.488407
##              9       10
## AIC(n) 1.476764 1.534471
## HQ(n)  2.108092 2.233441
## SC(n)  3.032453 3.256841
## FPE(n) 4.419438 4.697686
```

```
##                ME     RMSE      MAE      MPE     MAPE      ACF1 Theil's U
## Test set 77.88705 98.27504 77.91749 21.68205 21.69732 0.9433964  5.042171
```

```
##
##  Portmanteau Test (asymptotic)
##
## data:  Residuals of VAR object var1
## Chi-squared = 113.04, df = 81, p-value = 0.01081
```

```
##
##  Portmanteau Test (asymptotic)
##
## data:  Residuals of VAR object var2
## Chi-squared = 104.1, df = 72, p-value = 0.007962
```

```
##
##  Portmanteau Test (asymptotic)
##
## data:  Residuals of VAR object var3
## Chi-squared = 104.1, df = 72, p-value = 0.007962
```

SECTION B: Mathematical Questions

Question B1 (5 points): What is the unconditional mean of an AR process Y(t) = 10 + 1.1Y(t-1) + e(t)? unconditional mean = c/(1-phi) which is 10/(1-1.1), here, the coefficient phi is more than 1, hence the unconditional mean is indefinite and the series explodes. Phi has to be between 0 and 1 for there to be an unconditional mean. Question B2 (5 points): What is the unconditional mean of an MA process Y(t) = 10 + 20e(t-1) + 50e(t-2) - 100e(t-3) + e(t)? The unconditional mean of an MA process is 0. Question B3 (5 points): In an ARIMA-X model, residuals from the first stage regression are stationary [True or False]? Residuals from the second stage modelling are stationary [True or False]? False, regression errors are serially correlated, they could exhibit non-stationarity. True, innovation errors are white noise, they are stationary (zero mean, constant variance).

Section C: Short answer questions

Question C1 (13 points): When would you use a Vector Auto Regression (VAR) model over ARIMA-X, and vice versa? Discuss the similarities, differences, pros and cons of each approach I would use a VAR over ARIMA-x when the forecasts are long into the future. VAR is more accurate in long-term forecasting with extensive data, while ARIMA-x is more accurate in short term forecasting with less data. The reason being VAR is more prone to overfitting, having coefficients that scale quickly (k + pkˆ2). Recall that each coefficient needs 10 observations, hence VAR is only useful if we have a lot of data, perhaps in quarterly or monthly as opposed to annual data. ARIMA-x is less prone to overfitting, while ARIMA-x is more accurate short term as the exogenous variables also need to be forecasted. In the long-term, forecasting 2 things at once (exogenous and dependant variable of interest) will definitely lead to poor results.

Similarities: Both methods are considered multivariate forecasting, which means we are forecasting multiple things at once. Both methods involve other independant variables in forecasting, as opposed to just unvariate modelling (ARIMA/Exponential smoothing/ etc. . .) framework of using past values/long-term level of the same dependant variable in forecasting processes. Hence, there are patterns and relationships that ARIMA-x and VAR are able to capture and account for in model processes that a usual ARIMA would not, thus justifying why we should use these methods.

Differences: VAR and ARIMA-x differs in terms of how the series of the dependant variable we are interested in relates to the other independent time series used to forecast the dependant series.

I would use a VAR when variables are interacting with one another, interaction variables, in a bi-directional manner. An example using simple macroeconomics is consumption and income. Higher consumption leads to higher income, higher income leads to higher consumption. In this case, recalling the equation behind VAR, the lagged values of both income and consumption would have an effect on future values of both income and consumption. This is a result of the bi-directional relationship, which justifies the use of a VAR.

On the other hand, ARIMA-x is used when there are independant variables that affect the dependant variable (the one we are interested in), in a one directional relationship. An example would be climate change, studies have demonstrated that c02 emissions affect temperature, but temperature does not affect c02. In this case, C02 will be exogenous variable used to forecast temperature.

Another difference is that VAR forecasts multiple things at once, while ARIMA-x we forecast the exogenous variable first. Then, we use this forecasted exogenous variable values (for example in the “exog” flag of the forecast/arima function) to predict future values for our dependant variable of interest.

The pro of both models is that using other variables to forecast or explain the dependant variable of interests gives a more solidified and well-rounded statistical approach as opposed to univariate modelling using past values of dependant variable only. This is beacause it considers other factors and captures statistical relationships or patterns that might affect the dependant variable, and is more insightful than simply using old values or long term trend of just the dependant variables in forecasting.

The pros of VAR is that it provides stong long-term forecasts. The con is that it is prone to overfitting and cannot be used when data is limited, and that we have to ensure variables are granger causing one another. Also, we have to ensure that they past the serial.test(). If fails, add more lags until it ideally passes.

The pro of ARIMA-x is that it provides strong short-term forecasts. Additionally, it is less prone to overfitting. The con is that long-term forecasts are weaker. Adding on to this point, we need highly accurate forecasts of exogenous variables in the future. If these exogenous variables are forecasted poorly, then using it to forecast our dependant variable of interest will definitely lead to more inaccurate results.

Question C2 (4 points): What are the 2 qualitative features of the GARCH model in volatility forecasting that result in more realistic forecasts? 1. Mean reverting property of volitality. 2. Volitality clustering

Question C3 (5 points): Discuss the use of ex-ante and ex-post forecasts in “debugging” the source of inaccuracies in ARIMA-X based forecasts ex-ante forecasts refer to forecasts using observations we have already observed to evaluate model performance, while ex-post forecasts refer to forecasts into the future in which we have no observations of future values to evaluate the effectiveness of our models. When it comes to debugging the source of inaccuracies, I believe we use ex-ante forecasts. We do so by checking the out-of-sample performance to see it it is satisfactory (RMSE, MAPE that are decent), and the ljung box test on trained ARIMA-x model to see if there is any time series information still present in the residuals. If there is, build another model, until it passes the Ljung box test. Additionally, we could consider and build multiple ARIMA models, and use the AICc value to help us determine the best model in terms of preventing overfitting.

Question C4 (4 points): When should you use multiplicative versus additive time series decomposition? How do you convert a multiplicative time series to additive? We use multiplicative series decomposition when the trend is exponentially increasing, and additive decomposition when trend is increasing at a constant rate. We convert multiplicative time series to additive by applying ln, such that they can be added together as a property of the ln function.

Question C5 (5 points): What is the difference between covariance stationary and stationary? Covariance stationary means that the first and second moments , mean and variance are stationary. Stationary means that all moments, including mean and variance but also including kertosis (not sure how to spell) and skewness are stationary.

Question C6 (4 points): What do we (usually) need to make a variable stationary before modelling? We have to apply nsdiffs and ndiffs. Then apply differencing as needed (check if seasonal differencing is needed first, and perform seasonal differencing. Then check if non-seasonal differencing is needed, then non-seasonal differencing). Then we can either autoplot to visually investigate stationarity, or even better, use statistical tests by using the kpss unit root test, and see if we accept the null (series is stationary). Additionally, we could also, in the pre-processing steps, apply a boxcox transformation, which aims to make variance as constant as possible.

Section D: Design questions

Question D1 (10 points): You have two data series histories, which are cointegrated. Discuss the following points in any order you wish: (i) what does cointegration mean? (ii) what is the 'best' model you can build to predict both prices into the future, and why (iii) how does this model make use of the feature that both price series are cointegrated? i. Cointegration means that while both time series are non stationary l(1) etc, the linear combination of them are stationary. Hence, the difference between both series are mean reverting, constant variance and exhibit stationarity. ii. The best model to predict both prices in the future is the VECM model, Vector error correction model, which is an extension of the VAR. It adds onto the VAR model, which already contains past lagged values of multiple variables, by adding an ECM term, which will help to make up for deviations away from the long term level. iii. The feature that is used by exploiting cointegration is the ECM, whereby there is a mean reverting property of the deviations from long-term difference. I do not have the exact formula of VECM at hand, but I recall that in the RHS and second component (ECM component) of the equation, there is a tuning parameter that will tune the deviations from the expected long-term deviance accordingly. To elaborate, if the difference in prices of 2 cointegrated goods (eg. pepsi and cola), were to increase, then the parameter will decrease, in order to make up for the increase in the difference of prices between pepsi and cola. The key here is that cointegration means that the difference between the 2 prices have to be stationary and there are mean reverting properties, and hence the error correction component will be able to ensure this property.