EAS509 Final Exam

Submit your answers as a single pdf attach all R code. Failure to do so will result in grade reduction.

The exam must be done individually, with no discussion or help with others. Breaking this rule will result in an automatic 0 grade.

Part A (30 points) - each question worth 1 points

Some questions have multiple answers

1. Which simple forecasting method says the forecast is equal to the mean of the historical data?

a. Average Method
b. Naïve Method
c. Seasonal Naïve Method
d. Drift Method

Answer:a. Average Method

2. Which simple forecasting method says the forecast is equal to the last observed value?

a. Average Method
b. Naïve Method
c. Seasonal Naïve Method
d. Drift Method

Answer:b. Naïve Method

3. Which simple forecasting method is equivalent to extrapolating a line draw between the first and lost observations?

a. Average Method
b. Naïve Method
c. Seasonal Naïve Method
d. Drift Method

Answer:d. Drift Method

4. Which of the following is an assumption made about forecasting residuals during point forecast?

a. Residuals are normally distributed
b. Residuals are uncorrelated
c. Residuals have constant variance

d. None of the above

Answer: b. Residuals are uncorrelated

5. Which of the following is an assumption made about forecasting residuals during interval forecasting? (multiple answers)

a. Residuals have mean zero
b. Residuals are normally distributed
c. Residuals have constant variance
d. None of the above

Answer:a. Residuals have mean zero,b. Residuals are normally distributed & c. Residuals have constant variance all should present for full score

6. What is the consequence of forecasting residuals that are not uncorrelated?

a. Prediction intervals are difficult to calculate
b. Information is left in the residuals that should be used
c. Forecasts are biased
d. None of the above

Answer: b. Information is left in the residuals that should be used

7. What is the consequence of forecasting residuals that don't have mean zero?

a. Prediction intervals are difficult to calculate
b. Information is left in the residuals that should be used
c. Forecasts are biased
d. None of the above

Answer: c. Forecasts are biased

8. Which measure of forecast accuracy is scale independent?

a. MAE
b. MSE
c. RMSE
d. MAPE

Answer: d. MAPE

9. Calculation of forecasts is based on what?

a. Test set
b. Training set
c. Both
d. Neither

Answer: b. Training set

10. Forecast accuracy is based on what?

a. Test set
b. Training set
c. Both
d. Neeither

Answer: a. Test set

11. A series that is influenced by seasonal factors is known as what?

a. Trend
b. Seasonal
c. Cyclical
d. White Noise

Answer: b. Seasonal

12. Data that exhibits rises and falls that are not of a fixed period is known as what?

a. Trend
b. Seasonal
c. Cyclical
d. White Noise

Answer: a. Trend & c. Cyclicale either or all is ok for full credit

13. Data that is uncorrelated over time is known as what?

a. Trend
b. Seasonal
c. Cyclical
d. White Noise

Answer: d. White Noise

14. Which of the following time series decomposition models is appropriate when the magnitude of the seasonal fluctuations are not proportional to the level?

a. Additive
b. Multiplicative
c. Both
d. Neither

Answer: a. Additive

15. Which of the following time series decomposition models is appropriate when the magnitude of the seasonal fluctuations are proportional to the level?

a. Additive

b. Multiplicative
c. Both
d. Neither

Answer: b. Multiplicative

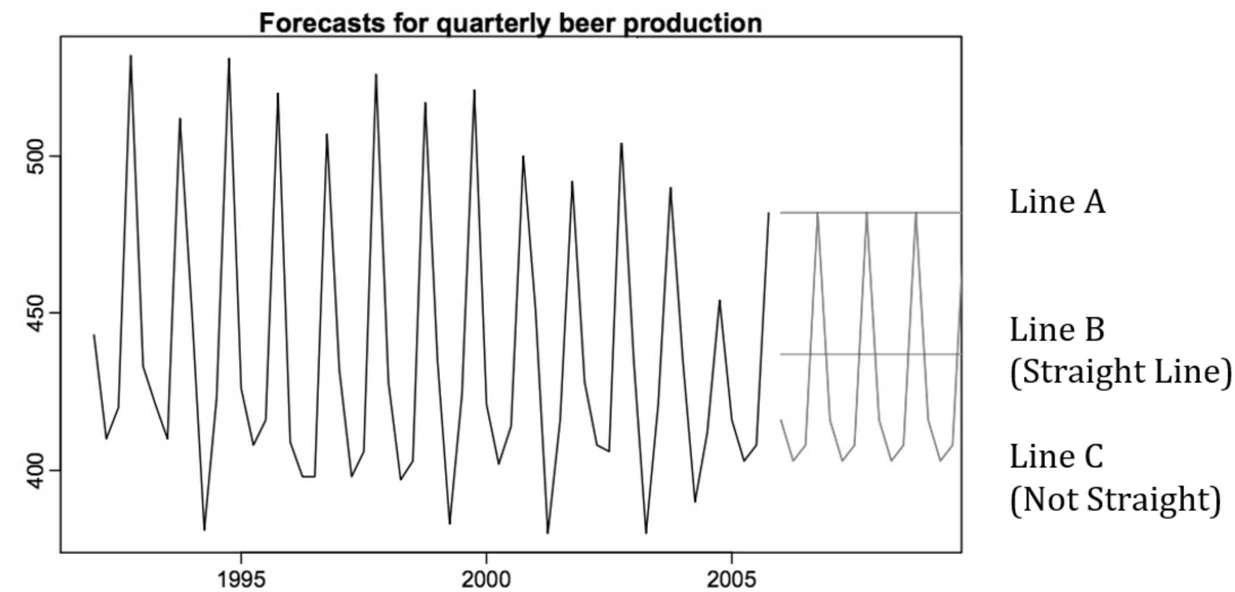


Exhibit 1

16. Refer to Exhibit 1. Line A is which simple forecasting method?CHECK

a. Average Method
b. Naïve Method
c. Seasonal Naïve Method
d. Drift

Answer: b. Naïve Method

17. Refer to Exhibit 1. Line B is which simple forecasting method?cHECK

a. Average Method
b. Naïve Method
c. Seasonal Naïve Method
d. Drift Method

Answer: a. Average Method

18. Refer to Exhibit 1. Line C is which simple forecasting method?CHECK

a. Average Method
b. Naïve Method
c. Seasonal Naïve Method
d. Drift Method

Answer: c. Seasonal Naïve Method

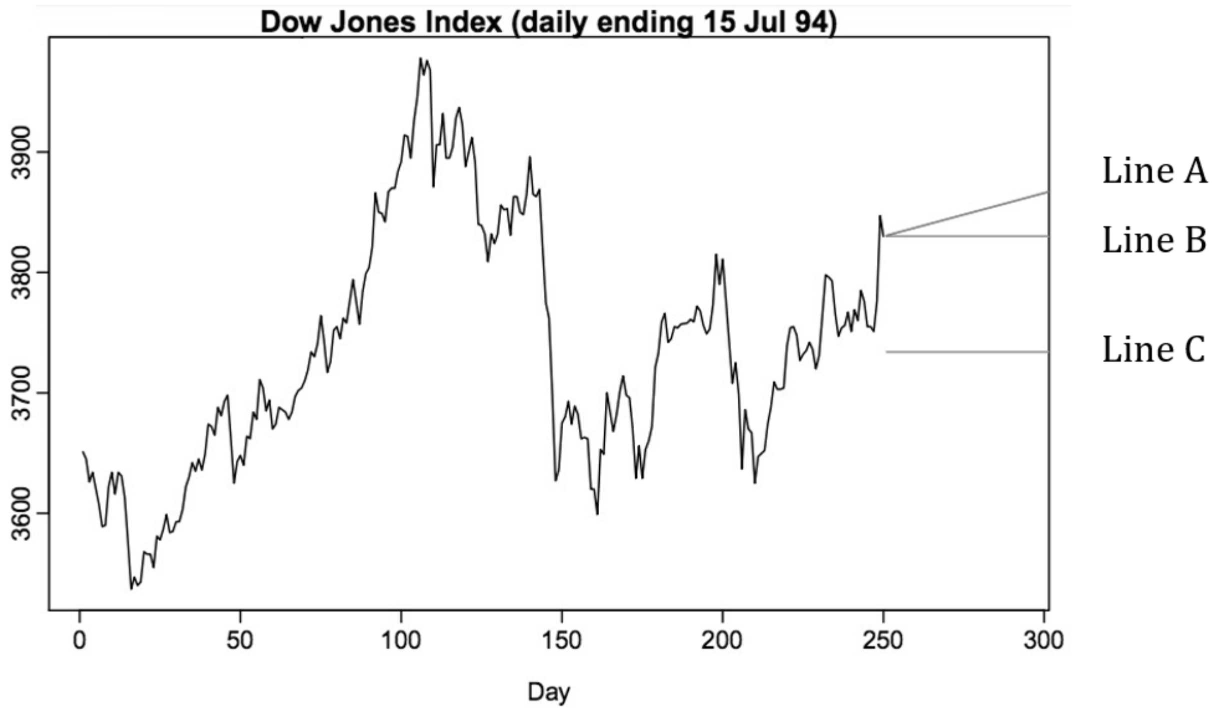


Exhibit 2

19. Refer to Exhibit 2. Line A is which simple forecasting method?

   a. Average Method
   b. Naïve Method
   c. Seasonal Naïve Method
   d. Drift Method

Answer: d. Drift Method

20. Refer to Exhibit 2. Line B is which simple forecasting method?

   a. Average Method
   b. Naïve Method
   c. Seasonal Naïve Method
   d. Drift Method

Answer: b. Naïve Method

21. Refer to Exhibit 2. Line C is which simple forecasting method?

   a. Average Method
   b. Naïve Method
   c. Seasonal Naïve Method
   d. Drift Method

Answer: a. Average Method

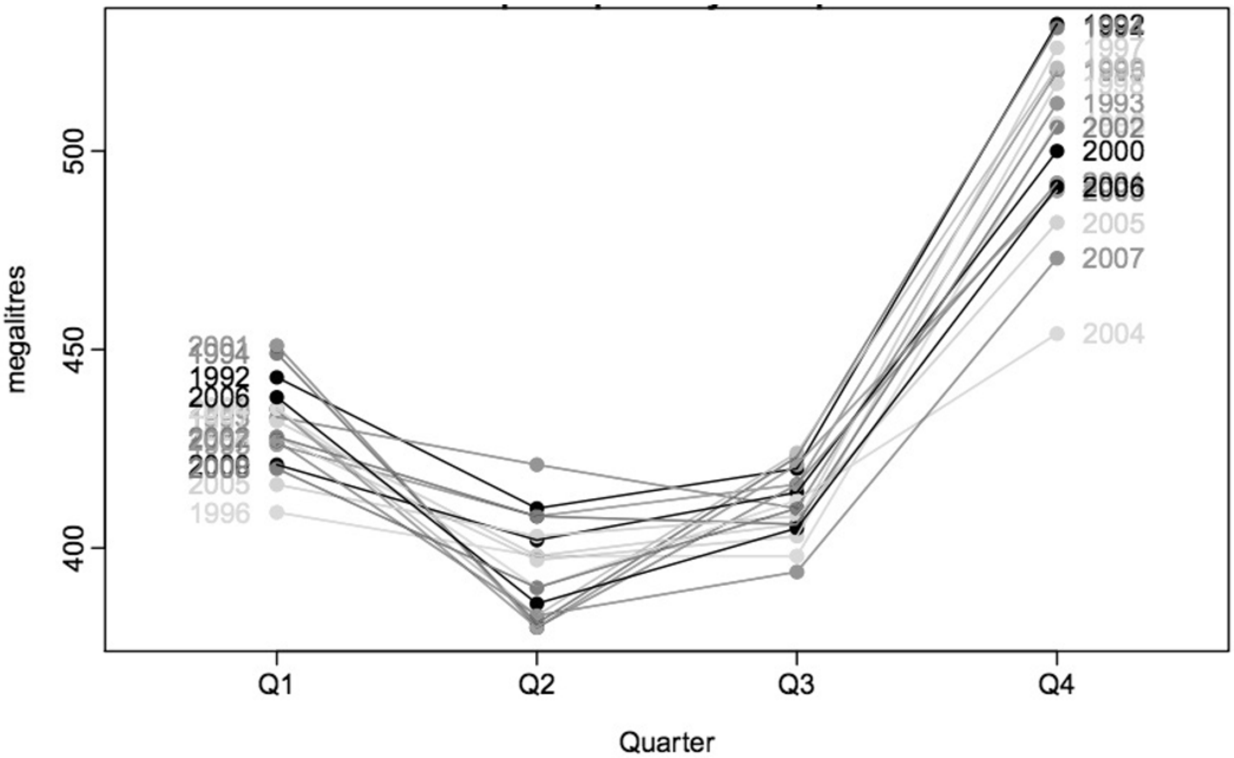


Exhibit 3

22. Refer to Exhibit 3. The peaks are in which quarter?

   a. Quarter 1
   b. Quarter 2
   c. Quarter 3
   d. Quarter 4

Answer: d. Quarter 4

23. Refer to Exhibit 3. The trough are in which quarter?

   a. Quarter 1
   b. Quarter 2
   c. Quarter 3
   d. Quarter 4

Answer: b. Quarter 2 there are few in Q3 but largly it is Q2

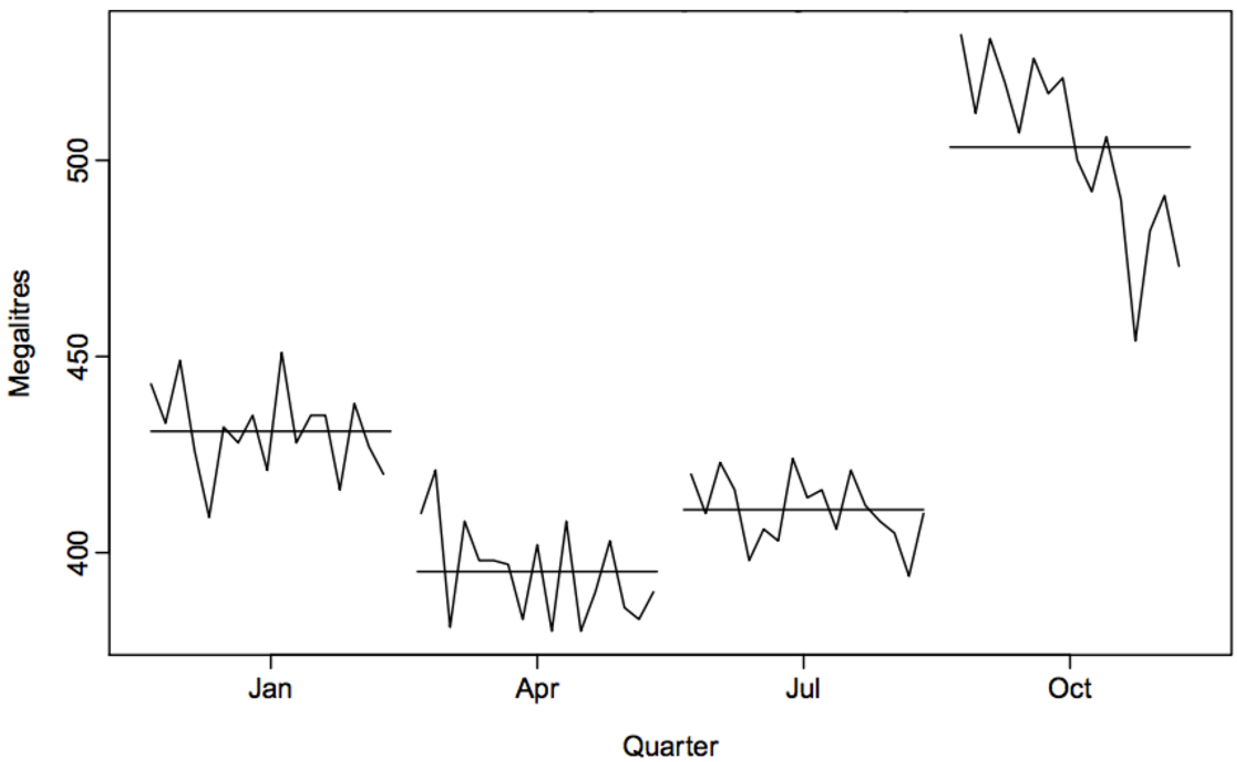


Exhibit 4

24. Refer to Exhibit 4. The peaks are in which quarter?

    a. Quarter 1
    b. Quarter 2
    c. Quarter 3
    d. Quarter 4

Answer: d. Quarter 4

25. Refer to Exhibit 4. The trough are in which quarter?

    a. Quarter 1
    b. Quarter 2
    c. Quarter 3
    d. Quarter 4

Answer: b. Quarter 2

26. Refer to Exhibit 4. In which quarter is there a decline in the seasonal affect?

    a. Quarter 1
    b. Quarter 2
    c. Quarter 3
    d. Quarter 4

Answer: d. Quarter 4

Figure 5

Year 1				Year 2			
Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
10	6	8	12	11	7	9	13

27. Refer to Figure 5. Using the average method, what is the forecast of Quarter 2 of Year 3? (Don't use a calculator.)

a. 7
b. 9.5
c. 13.85
d. 13

Answer: b. 9.5

28. Refer to Figure 5. Using the naïve method, what is the forecast of Quarter 2 of Year 3? (Don't use a calculator.)

a. 7
b. 9.5
c. 13.85
d. 13

Answer:d. 13

29. Refer to Figure 5. Using the seasonal naïve method, what is the forecast of Quarter 2 of Year 3? (Don't use a calculator.)

a. 7
b. 9.5
c. 13.85
d. 13
e. 7 **Answer: a. 7**

30. Refer to Figure 5. Using the drift method, what is the forecast of Quarter 2 of Year 3? (Don't use a calculator.)

a. 7
b. 9.5
c. 13.85
d. 13

Answer: c. 13.85

Part B (30 points)

Choose a series from us_employment.cvs, the total employment in leisure and hospitality industry in the United States (see, title column).

a. Produce an STL decomposition of the data and describe the trend and seasonality. (4 points)

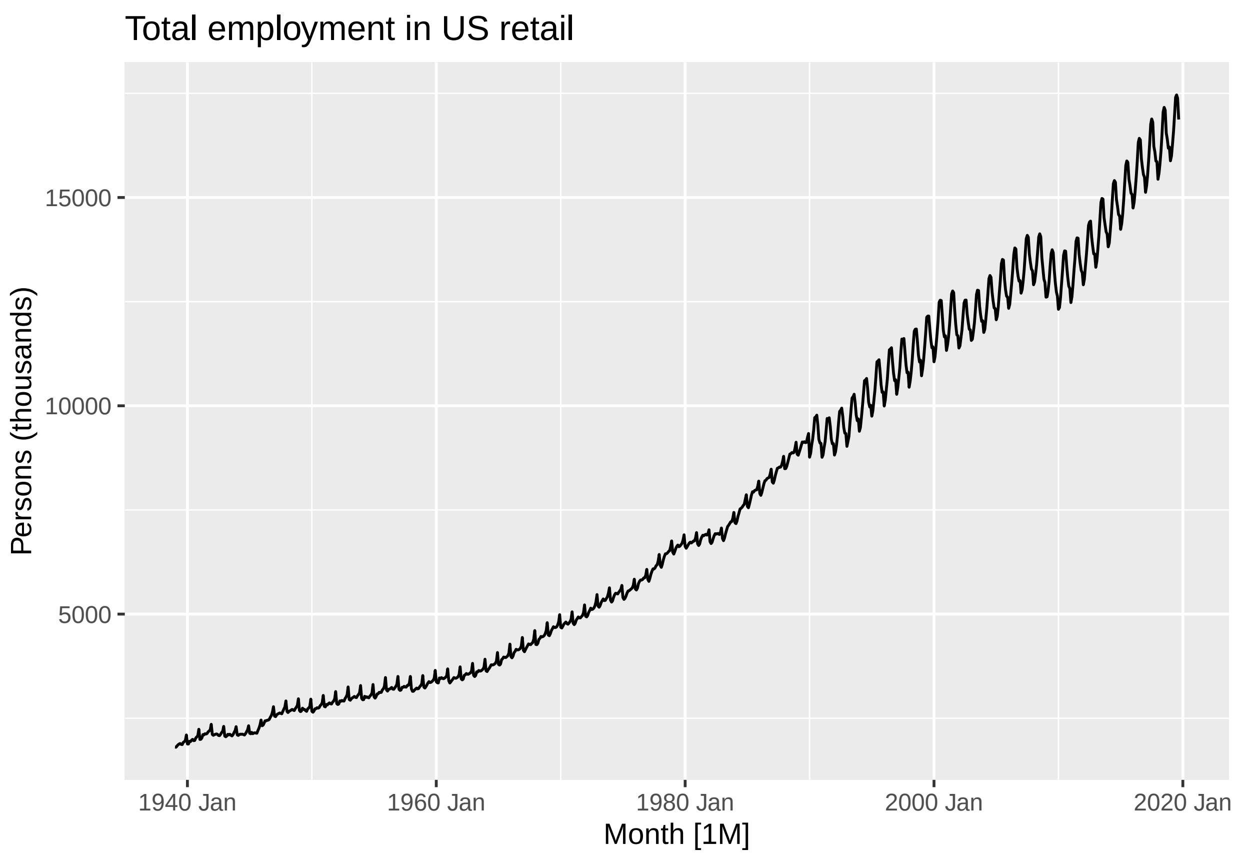
```
# I am reading the data of Lesiure and hospitality industry
data <- fread('us_employment.cvs')
us_employemnt <- data[Title == 'Leisure and Hospitality']
head(us_employemnt)
```

```
##       Month     Series_ID                   Title Employed
## 1: 1939 Jan CEU7000000001 Leisure and Hospitality     1807
## 2: 1939 Feb CEU7000000001 Leisure and Hospitality     1804
## 3: 1939 Mar CEU7000000001 Leisure and Hospitality     1834
## 4: 1939 Apr CEU7000000001 Leisure and Hospitality     1863
## 5: 1939 May CEU7000000001 Leisure and Hospitality     1882
## 6: 1939 Jun CEU7000000001 Leisure and Hospitality     1894
```

```
#Now lets convert the data into timeseries.

us_employemnt_TS <- us_employemnt %>%
  select(Month, Employed) %>%
  mutate(Month = yearmonth(Month)) %>%
  as_tsibble(index = Month)

head(us_employemnt_TS)
```

```
## # A tsibble: 6 x 2 [1M]
##      Month Employed
##      <mth>    <dbl>
## 1 1939 Jan     1807
## 2 1939 Feb     1804
## 3 1939 Mar     1834
## 4 1939 Apr     1863
## 5 1939 May     1882
## 6 1939 Jun     1894
```

```
#Before Decomposition Plot
us_employemnt_TS %>%
  autoplot(Employed) +
  labs(
    y = "Persons (thousands)",
    title = "Total employment in US retail"
  )
```

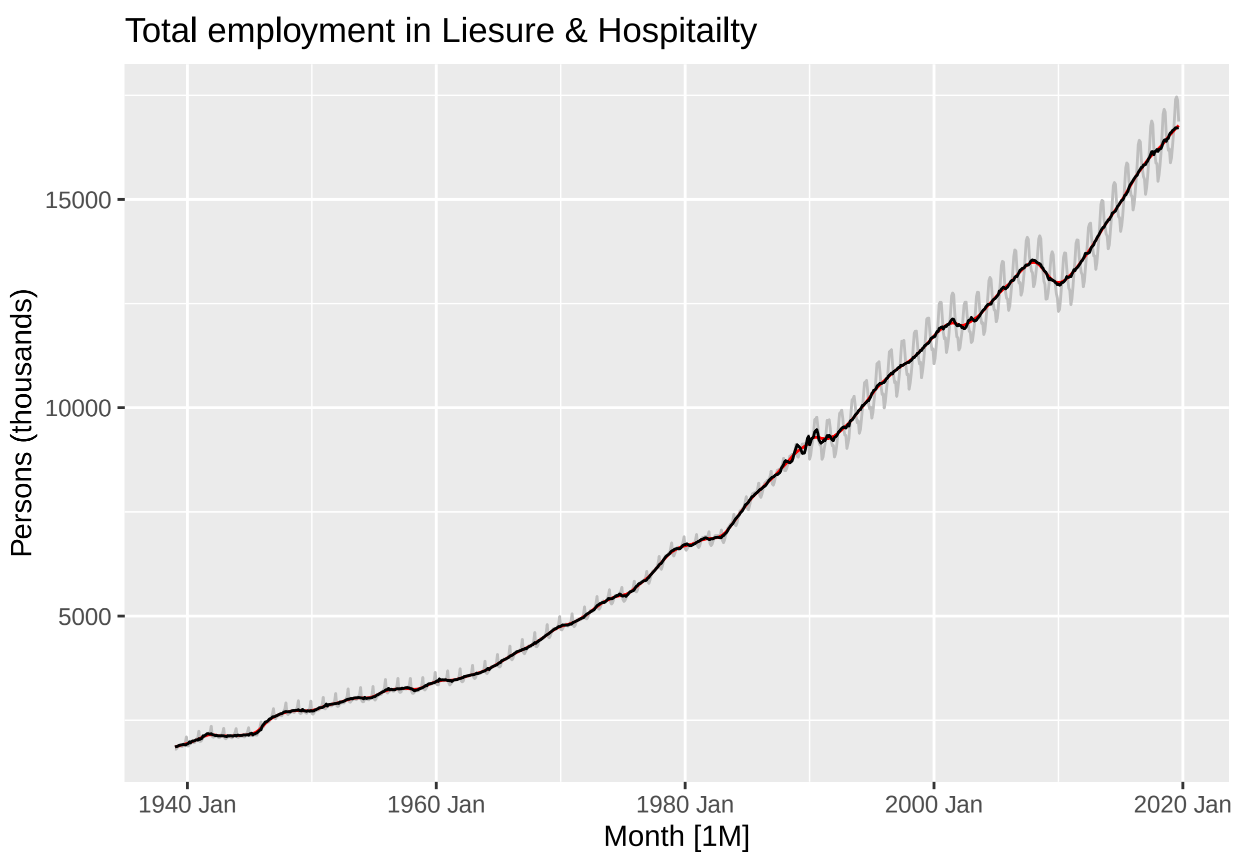



```
#STL Decomposition
us_emp_dcmp <- us_employemnt_TS %>%
  model(stl = STL(Employed))
components(us_emp_dcmp)
```

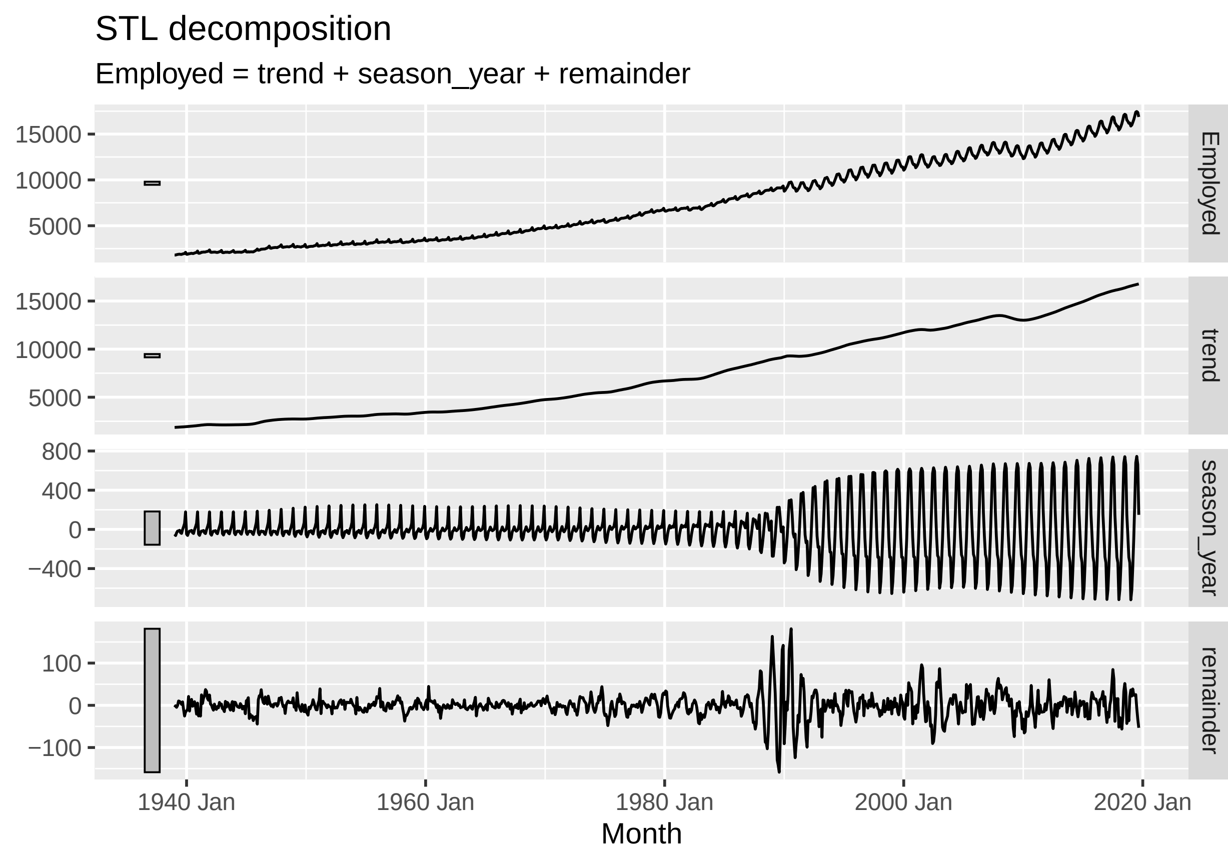
```
## # A dable: 969 x 7 [1M]
## # Key:      .model [1]
## # :         Employed = trend + season_year + remainder
##    .model    Month Employed trend season_year remainder season_adjust
##    <chr>     <mth>    <dbl> <dbl>       <dbl>     <dbl>         <dbl>
##  1 stl    1939 Jan     1807 1861.      -49.8     -4.55          1857.
##  2 stl    1939 Feb     1804 1868.      -62.8     -0.973         1867.
##  3 stl    1939 Mar     1834 1874.      -35.5     -4.77          1869.
##  4 stl    1939 Apr     1863 1881.      -17.2     -0.464         1880.
##  5 stl    1939 May     1882 1887.      -15.5     10.3           1898.
##  6 stl    1939 Jun     1894 1894.       -9.18     9.32          1903.
##  7 stl    1939 Jul     1873 1900.      -33.9      6.43          1907.
##  8 stl    1939 Aug     1868 1907.      -41.4      2.09          1909.
##  9 stl    1939 Sep     1920 1914.       -2.26     8.13          1922.
## 10 stl    1939 Oct     1938 1921.       26.6     -9.60          1911.
## # i 959 more rows
```

```
#After Decomposition Plot
us_employemnt_TS %>%
  autoplot(Employed, color = "gray") +
```

```r
  autolayer(components(us_emp_dcmp), trend, color = "red") +
  autolayer(components(us_emp_dcmp), season_adjust, color = "black") +
  labs(
    y = "Persons (thousands)",
    title = "Total employment in Liesure & Hospitailty"
  )
```




```r
components(us_emp_dcmp) %>% autoplot()
```

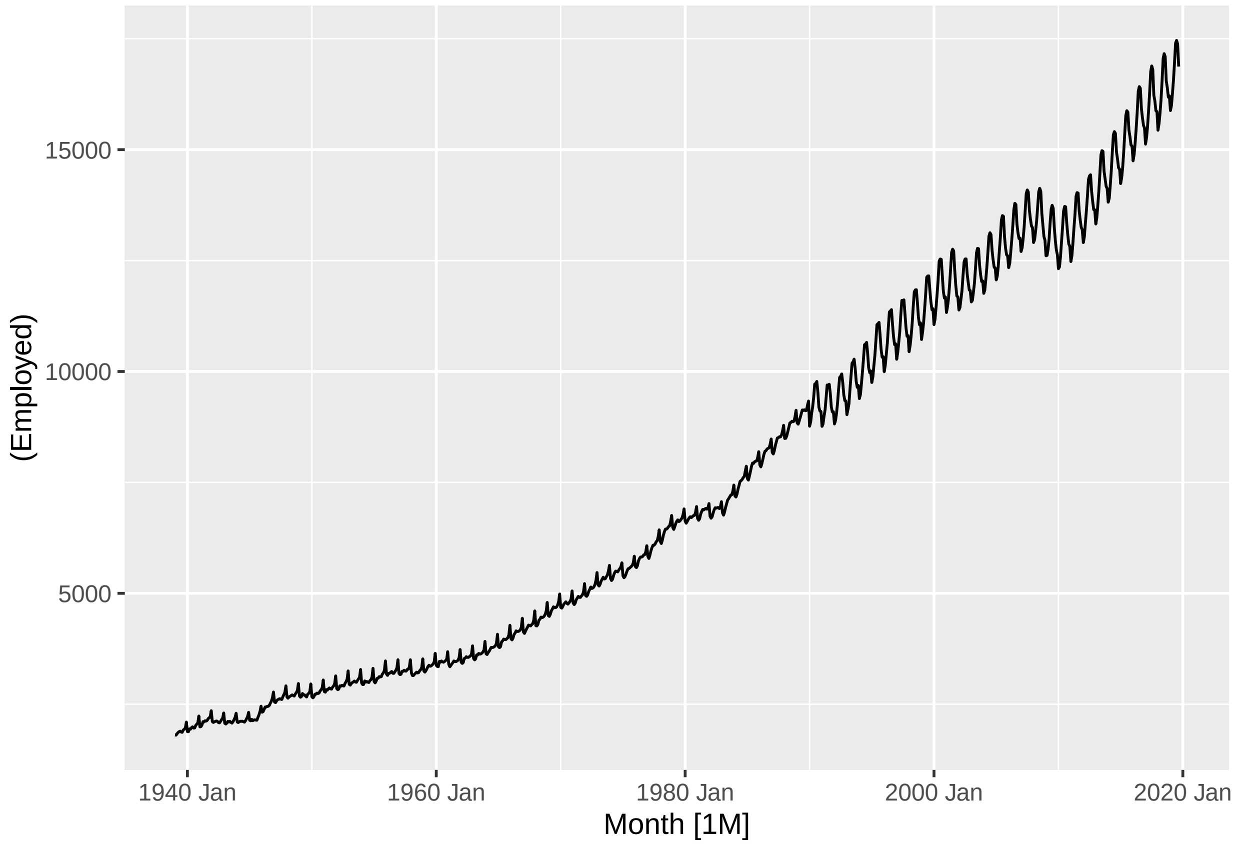


#The trend is increasing over years but we can see some drops in between 2002 and 2010. And in the initial years from 1940 to 1989, we cant see much variations in seasionality, but later especially in 1990 we can see the seasonnality is maximum. b. Do the data need transforming? If so, find a suitable transformation.(4 points)

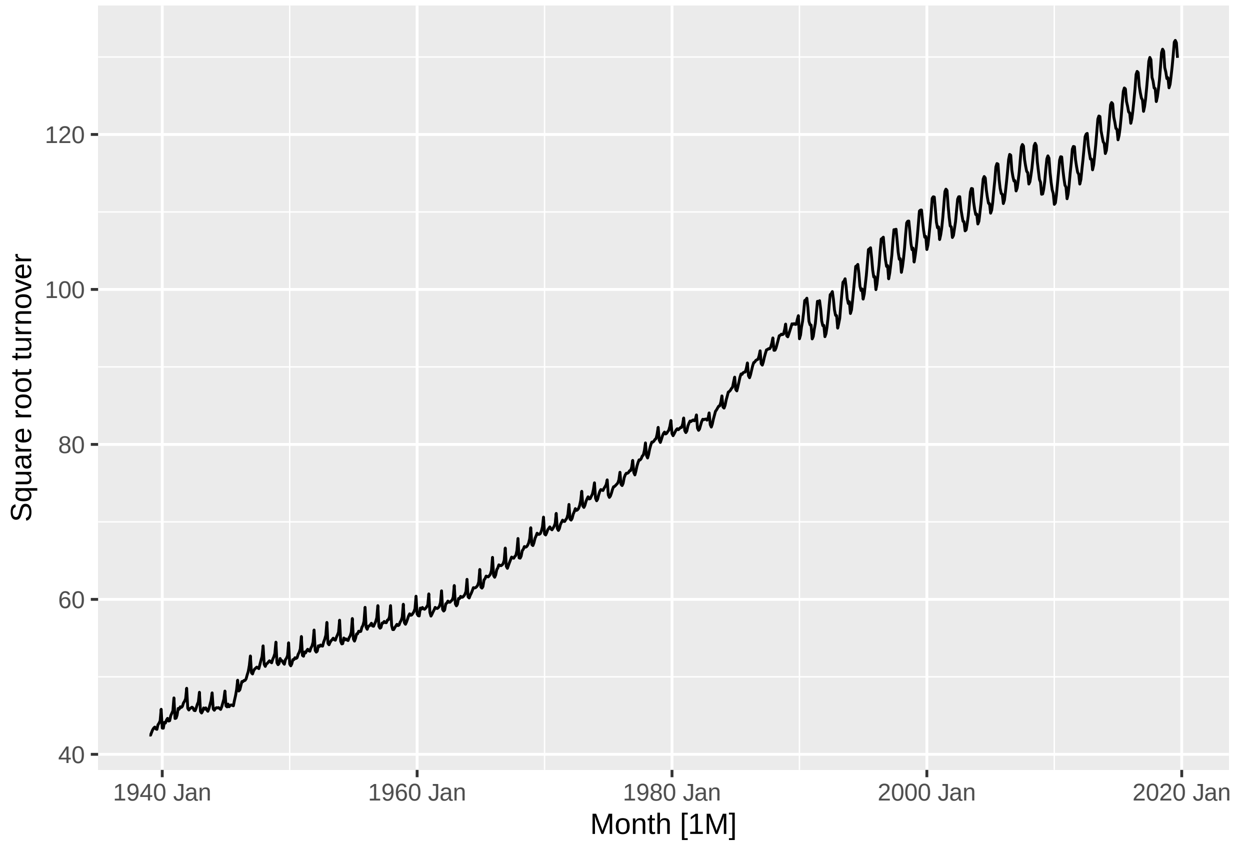
```
us_employemnt_TS %>%
  features(Employed, features = guerrero)
```

```
## # A tibble: 1 x 1
##   lambda_guerrero
##             <dbl>
## 1          -0.216
```

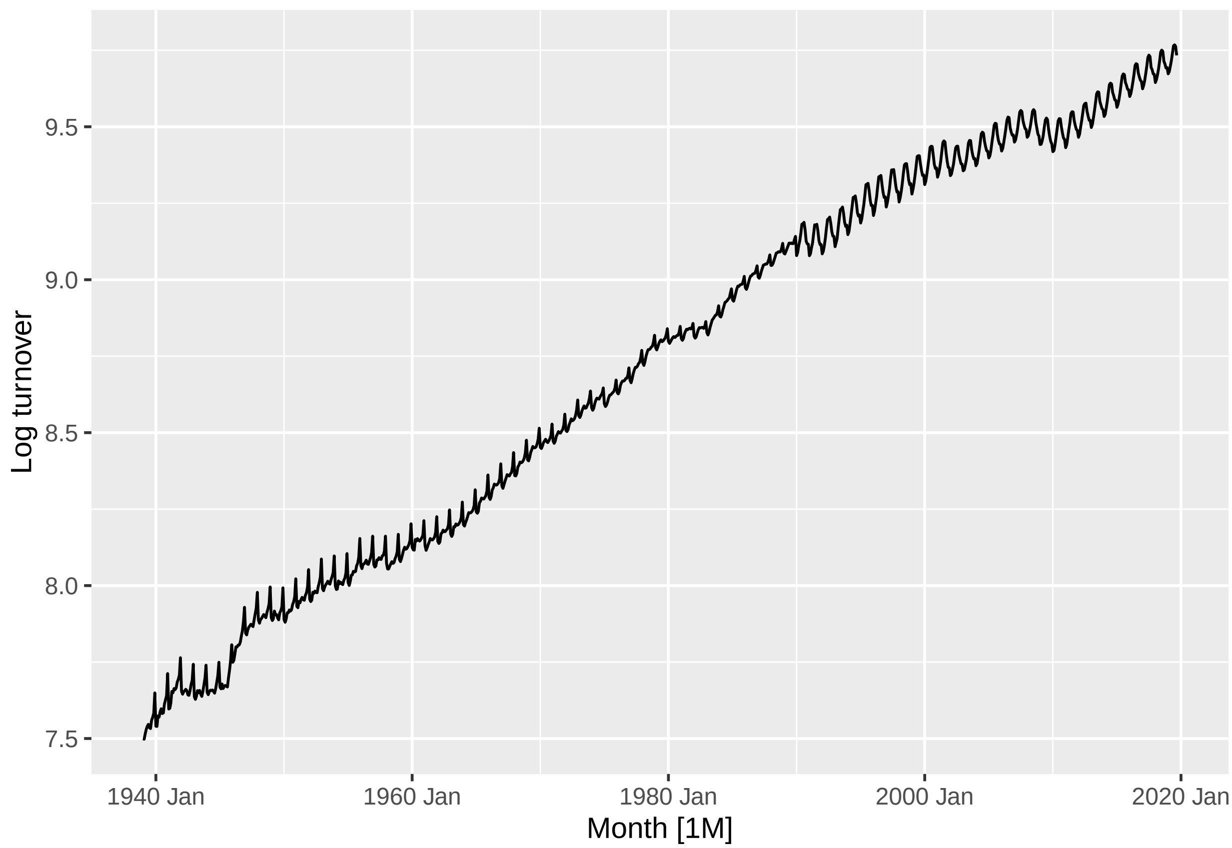
```
us_employemnt_TS %>% autoplot((Employed))
```



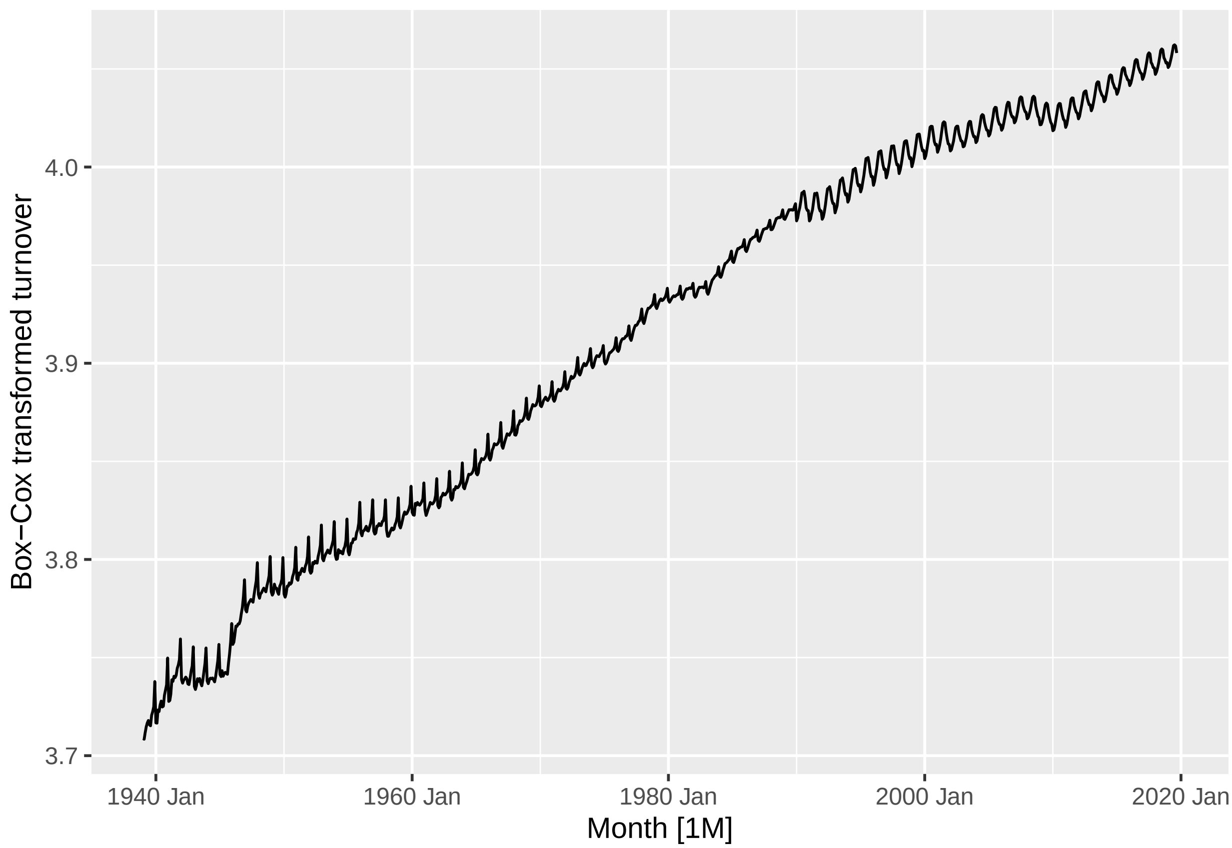

```
us_employemnt_TS %>% autoplot(sqrt(Employed)) +
  labs(y = "Square root turnover")
```




```r
us_employemnt_TS %>% autoplot(log(Employed)) +
  labs(y = "Log turnover")
```




```r
us_employemnt_TS %>% autoplot(box_cox(Employed,-0.2164477)) +
  labs(y = "Box-Cox transformed turnover")
```



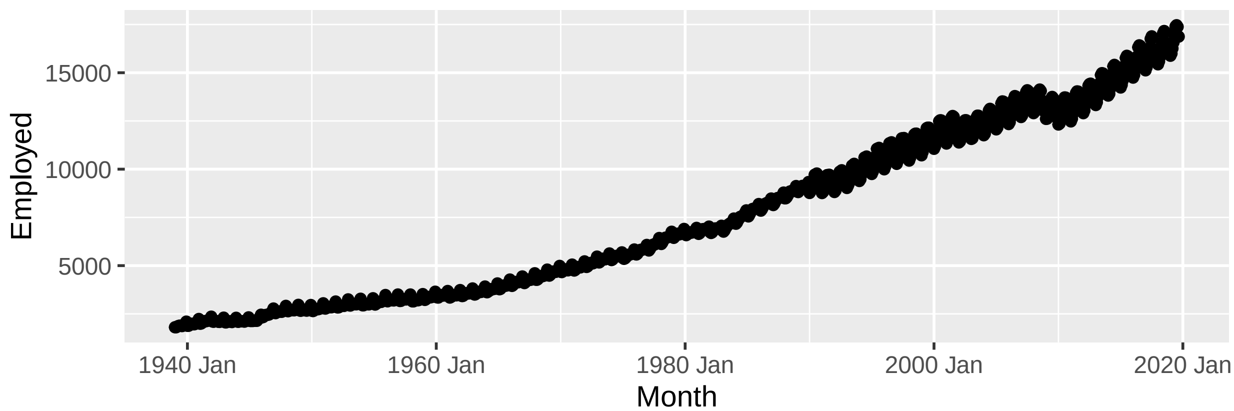
#Yes data needs transaformation.I could find that Box-cox and log are very similar but BoxCox is better, So I transfromed our data and stored into new variable.

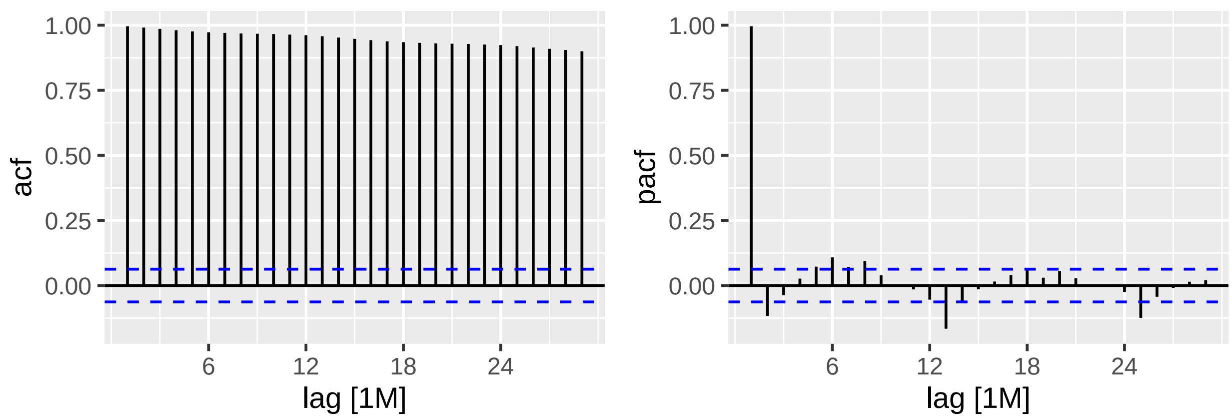
```r
us_employemnt_Trans <- us_employemnt_TS %>%
  mutate(Employed = box_cox(Employed, lambda = -0.2164477))

head(us_employemnt_Trans)
```

```
## # A tsibble: 6 x 2 [1M]
##      Month Employed
##      <mth>    <dbl>
## 1 1939 Jan     3.71
## 2 1939 Feb     3.71
## 3 1939 Mar     3.71
## 4 1939 Apr     3.71
## 5 1939 May     3.72
## 6 1939 Jun     3.72
```

c. Are the data stationary? If not, find an appropriate differencing which yields stationary data.(4 points)

```r
#ACF & PCF plots
gg_tsdisplay(us_employemnt_TS,Employed,plot_type='partial')
```






```
ndiffs(us_employemnt_TS$Employed,alpha=0.05)
```

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

As per my observations from PACF and ACF plots, I feel that data is not stationary. Also 1st differencing should be selected.

d. Identify a couple of ARIMA models that might be useful in describing the time series. Which of your models is the best according to their AICc values?(5 points)

```
fit <- us_employemnt_Trans %>%
  model(
    arima_auto = ARIMA(Employed),
    #from ACF&pacf plot
    arima1 = ARIMA(Employed~0+pdq(12,1,0)+PDQ(1,0,0)),
    arima2 = ARIMA(Employed~0+pdq(2,1,2)+PDQ(0,1,2)),
    arima3 = ARIMA(Employed~0+pdq(2,1,0))
  )
accuracy(fit)
```

```
## # A tibble: 4 x 10
##   .model     .type         ME     RMSE      MAE       MPE    MAPE   MASE RMSSE     ACF1
```

```
##   <chr>      <chr>     <dbl>   <dbl>   <dbl>    <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 arima_auto Train~ -1.56e-5 9.77e-4 5.73e-4 -4.09e-4 0.0148 0.126 0.172 0.0306
## 2 arima1     Train~  9.43e-7 9.12e-4 5.67e-4  3.13e-5 0.0147 0.124 0.161 0.00982
## 3 arima2     Train~ -1.70e-5 9.92e-4 5.76e-4 -4.45e-4 0.0149 0.126 0.175 0.0246
## 4 arima3     Train~ -1.67e-5 9.87e-4 5.83e-4 -4.37e-4 0.0151 0.128 0.174 0.00330
```

```r
report(fit[1])
```

```
## Series: Employed
## Model: ARIMA(1,1,2)(2,1,1)[12]
##
## Coefficients:
##          ar1      ma1     ma2     sar1     sar2    sma1
##       0.5861  -0.6248  0.2476  -1.0777  -0.4674  0.5957
## s.e.  0.0763   0.0755  0.0345   0.0668   0.0350  0.0699
##
## sigma^2 estimated as 9.746e-07:  log likelihood=5362.75
## AIC=-10711.5   AICc=-10711.39   BIC=-10677.47
```

```r
report(fit[2])
```

```
## Series: Employed
## Model: ARIMA(12,1,0)(1,0,0)[12]
##
## Coefficients:
##           ar1     ar2     ar3     ar4      ar5      ar6     ar7     ar8     ar9
##       -0.0811  0.1753  0.1469  0.0876  -0.0040  -0.0313  0.0020  0.0124  0.0157
## s.e.   0.0295  0.0295  0.0303  0.0306   0.0308   0.0308  0.0307  0.0308  0.0306
##         ar10    ar11     ar12    sar1
##       0.0234  0.1023  -0.4028  0.9927
## s.e.  0.0303  0.0296   0.0300  0.0026
##
## sigma^2 estimated as 8.447e-07:  log likelihood=5387.04
## AIC=-10746.08   AICc=-10745.64   BIC=-10677.83
```

```r
report(fit[3])
```

```
## Series: Employed
## Model: ARIMA(2,1,2)(0,1,2)[12]
##
## Coefficients:
##          ar1      ar2      ma1     ma2     sma1    sma2
##       0.6914  -0.1701  -0.7512  0.4173  -0.5077  0.0668
## s.e.  0.1376   0.1253   0.1281  0.1107   0.0327  0.0343
##
## sigma^2 estimated as 1.004e-06:  log likelihood=5345.93
## AIC=-10677.85   AICc=-10677.73   BIC=-10643.81
```

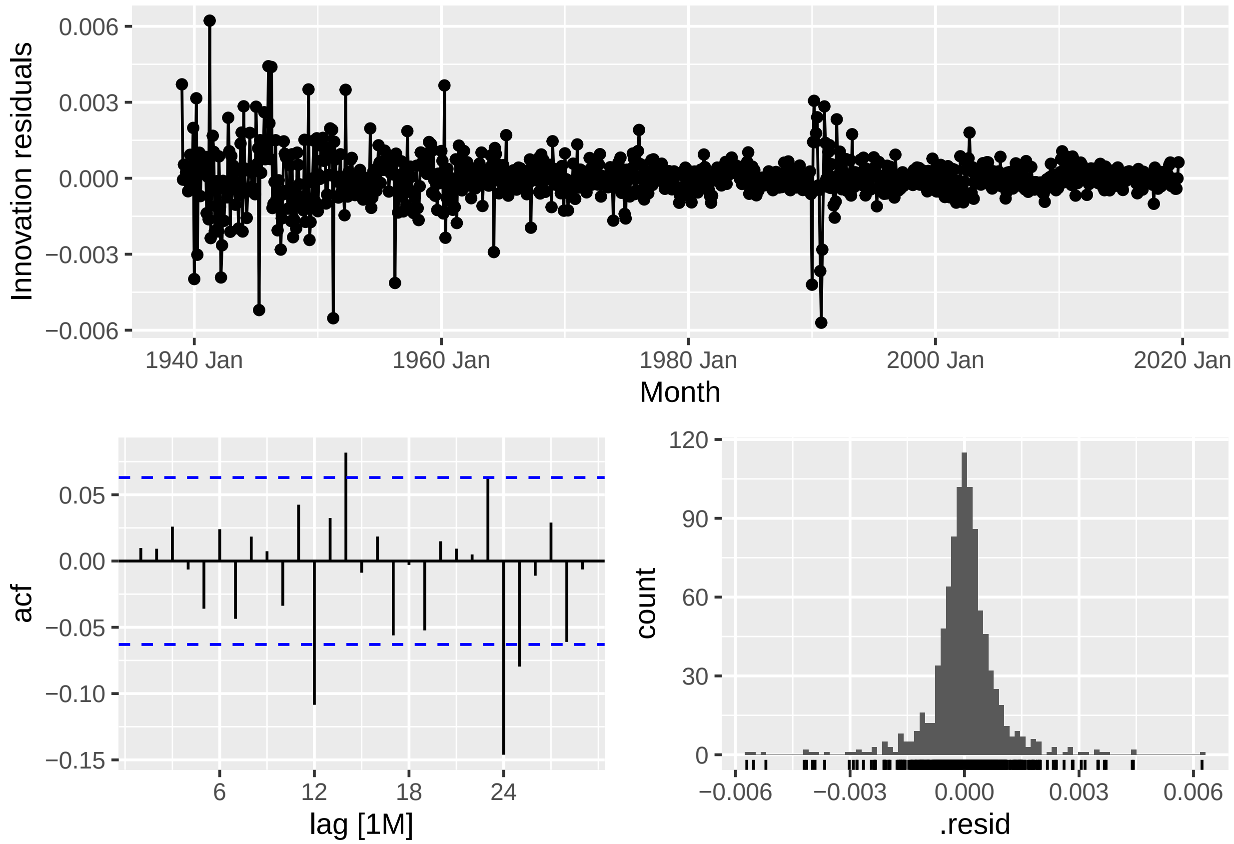
#So Ideally we will conslude that whichever AICc values is lower that should be best model. In our case arima1 has AIC (AIC=-10745.64) which is lowest among others, so we will say this model is best.

e. Estimate the parameters of your best model and do diagnostic testing on the residuals. Do the residuals resemble white noise? If not, try to find another ARIMA model which fits better.(5 points)

```
final_model <- fit %>% select(arima1)
report(final_model)
```

```
## Series: Employed
## Model: ARIMA(12,1,0)(1,0,0)[12]
##
## Coefficients:
##            ar1     ar2     ar3     ar4      ar5      ar6     ar7     ar8     ar9
##        -0.0811  0.1753  0.1469  0.0876  -0.0040  -0.0313  0.0020  0.0124  0.0157
## s.e.    0.0295  0.0295  0.0303  0.0306   0.0308   0.0308  0.0307  0.0308  0.0306
##           ar10    ar11     ar12    sar1
##         0.0234  0.1023  -0.4028  0.9927
## s.e.    0.0303  0.0296   0.0300  0.0026
##
## sigma^2 estimated as 8.447e-07:  log likelihood=5387.04
## AIC=-10746.08   AICc=-10745.64   BIC=-10677.83
```

```
gg_tsresiduals(final_model )
```



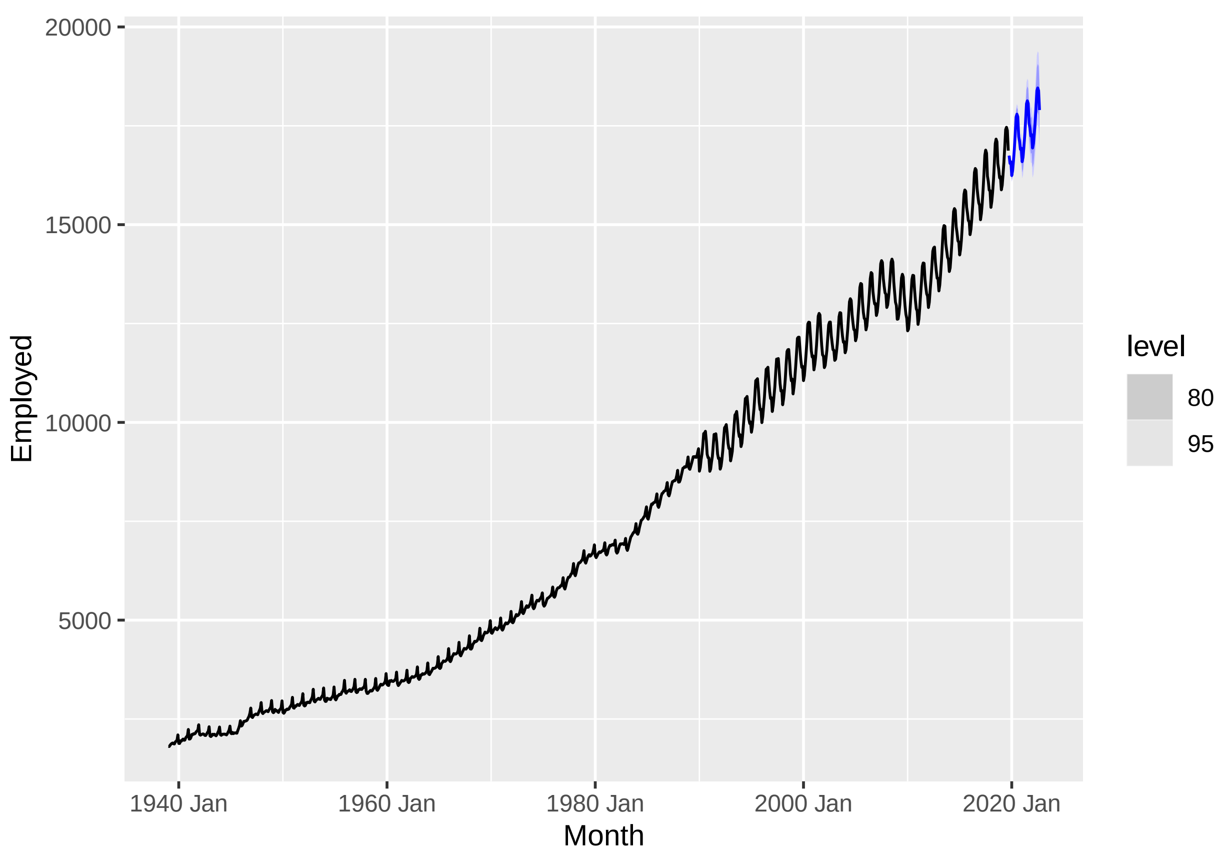

```
augment(final_model) %>% features(.innov, ljung_box, lag=10, dof=2)
```

```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 arima1    6.06     0.641
```

#We found that P values is high than 0.05 so I believe we have white noise and data with white noise is consistent.

f. Forecast the next 3 years of data. Get the latest figures from https://fred.stlouisfed.org/categories/11 to check the accuracy of your forecasts. (5 points)

```
arimafitt <- us_employemnt_TS%>%
model(
arima1 = ARIMA(Employed~0+pdq(12,1,0)+PDQ(1,0,0), stepwise = FALSE,approximation = FALSE)
)

ff <- forecast(arimafitt, h = 36)
ff %>% autoplot(us_employemnt_TS)
```




```
#latest data
latest<- fread('CEU7000000001.csv')
head(latest)
```

```
##          DATE CEU7000000001
```

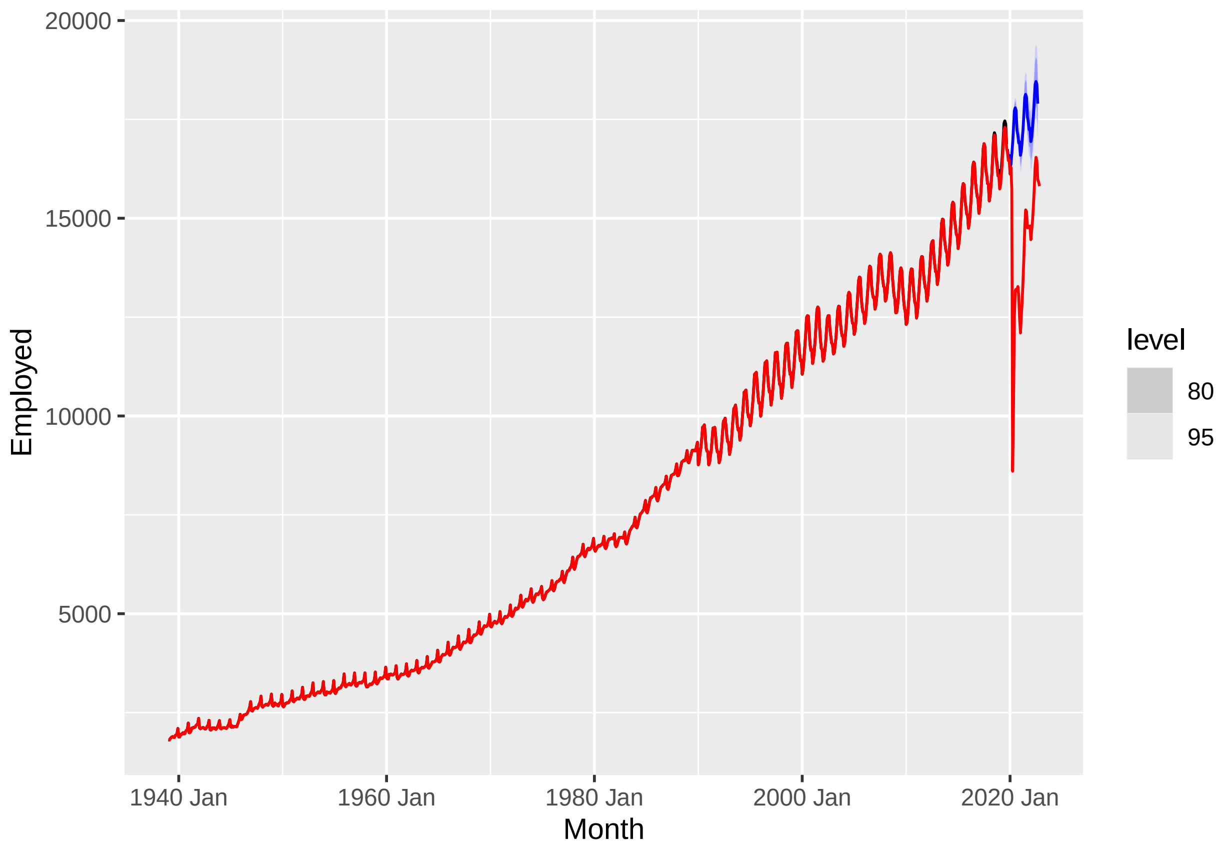
```
## 1: 1939-01-01          1807
## 2: 1939-02-01          1804
## 3: 1939-03-01          1834
## 4: 1939-04-01          1863
## 5: 1939-05-01          1882
## 6: 1939-06-01          1894
```

```
latest %>%
  mutate(DATE = yearmonth(DATE)) %>%
  tsibble(index = DATE) -> lateststts
lateststts
```

```
## # A tsibble: 1,007 x 2 [1M]
##        DATE CEU7000000001
##       <mth>         <int>
##  1 1939 Jan          1807
##  2 1939 Feb          1804
##  3 1939 Mar          1834
##  4 1939 Apr          1863
##  5 1939 May          1882
##  6 1939 Jun          1894
##  7 1939 Jul          1873
##  8 1939 Aug          1868
##  9 1939 Sep          1920
## 10 1939 Oct          1938
## # i 997 more rows
```

```
ff %>% autoplot(us_employemnt_TS)+
autolayer(lateststs, color = 'red')
```

```
## Plot variable not specified, automatically selected `.vars = CEU7000000001`
```



g. Eventually, the prediction intervals are so wide that the forecasts are not particularly useful. How many years of forecasts do you think are sufficiently accurate to be usable? (3 points)

```
#No I dont feel actual and predicted are accurate. It may be because of covid 19 which has impacted the
```

Part C (8 points)

Consider following transactions:

1. Eggs, Bread, Milk, Bananas, Onion, Yogurt
2. Dill, Eggs, Bread, Bananas, Onion, Yogurt
3. Apple, Eggs, Bread, Milk
4. Corn, Bread, Milk, Teddy Bear, Yogurt
5. Corn, Eggs, Ice Cream, Bread, Onion

a) Calculate by hand support, confidence and lift for following rules (without usage of apriory library, show your work)

- **{Bananas} -> {Yogurt} (2 points)**

```
N= 5
N_bananas =2
```

```
N_yogurt = 3
N_bananas_yogurt =2

support  = 2/5
confidence = 2/2

support_yogurt = 3/5

lift = 2/2 / 3/5 = 5/3 which is [confidence/support_yogurt]
```

- **{Corn, Bread}->{Onion} (3 points)**

```
N= 5
N_Onion =3
N_Corn = 2
N_corn_bread = 2
N_corn_bread_onion =1

support  = 1/5
confidence = 1/2

support_onion = 3/5

lift = 1/2 / 3/5 = 3/10 which is [confidence/support_onion]
```

- **{Bread}->{Milk, Yogurt} (3 points)**

```
N= 5
N_Bread = 5
N_Milk_Yogurt = 2
N_Bread_Milk_Yogurt = 2


support  = 2/5
confidence = 2/5

support_milk_yogurt = 2/5

lift = 2/5 / 2/5 = 1 which is [confidence/support_milk_yogurt]
```

Part D (32 points)

Online_Retail2.csv contains transaction from online store in long format (i.e. single item per line and lines with same InvoiceNo is single transaction).

a) Read data and convert it to transactions (hint: transactions function and format argument). (4 points)

```
order_data <- fread('Online_Retail2.csv')
order_data <- order_data %>%
  select(-CustomerID, -StockCode, -Country, -InvoiceDate)
head(order_data)
```

```
##    InvoiceNo                         Description Quantity UnitPrice
## 1:    536365  WHITE HANGING HEART T-LIGHT HOLDER        6      2.55
## 2:    536365                 WHITE METAL LANTERN        6      3.39
## 3:    536365      CREAM CUPID HEARTS COAT HANGER        8      2.75
## 4:    536365 KNITTED UNION FLAG HOT WATER BOTTLE        6      3.39
## 5:    536365      RED WOOLLY HOTTIE WHITE HEART.        6      3.39
## 6:    536365        SET 7 BABUSHKA NESTING BOXES        2      7.65
```

```r
order_trans <- transactions(order_data, format = 'long')
order_trans
```

```
## transactions in sparse format with
##  24446 transactions (rows) and
##  4211 items (columns)
```

```r
head(order_trans)
```

```
## transactions in sparse format with
##  6 transactions (rows) and
##  4211 items (columns)
```

b) Run summary on transactions. How many transactions are there? How many unique items? (4 points)

```r
summary(order_trans)
```

```
## transactions as itemMatrix in sparse format with
##  24446 rows (elements/itemsets/transactions) and
##  4211 columns (items) and a density of 0.005143444
##
## most frequent items:
## WHITE HANGING HEART T-LIGHT HOLDER           REGENCY CAKESTAND 3 TIER
##                               2302                               2169
##            JUMBO BAG RED RETROSPOT                      PARTY BUNTING
##                               2135                               1706
##            LUNCH BAG RED RETROSPOT                            (Other)
##                               1607                             519558
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 4440 1590 1080  812  791  671  654  634  635  562  568  505  513  537  555  557
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##  468  444  491  438  407  349  351  310  249  262  243  242  272  226  199  189
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
##  162  177  137  137  131  122  139  122  123  103   97  104  100   91   84   95
##   49   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64
##   88   86   57   65   78   70   73   50   65   51   36   61   40   29   43   39
##   65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80
##   39   42   34   40   29   33   39   23   25   34   26   21   19   27   15   13
##   81   82   83   84   85   86   87   88   89   90   91   92   93   94   95   96
##   20   21   15   23   17   17    9   17   11   12    9   15   16    7    5   10
```

```
##   97   98   99  100  101  102  103  104  105  106  107  108  109  110  111  112
##    9   13    5   11   11    3    6    9    2    4    7    4    4    4    7    3
##  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128
##    5    6    6    8    6    4    8    5    6   11    4    5    3    4    8    1
##  129  130  131  132  133  134  135  136  137  138  139  140  141  142  143  144
##    2    4    3    3    2    5    4    2    6    6    2    5    6    2    2    5
##  145  146  147  148  149  150  151  152  153  154  155  156  157  158  159  160
##    5    3    2    4    5    3    5    3    6    2    2    2    4    4    1    2
##  161  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176
##    3    3    3    2    5    4    1    4    4    2    2    4    3    4    2    5
##  177  178  179  180  181  182  183  184  185  186  187  188  189  190  191  192
##    5    4    2    4    2    6    4    3    3    3    2    3    4    4    2    3
##  193  194  195  196  197  198  199  202  203  204  205  206  207  208  210  211
##    2    3    3    4    2    2    3    2    5    5    1    2    1    4    1    4
##  212  213  214  215  216  217  218  219  220  222  223  224  225  226  227  228
##    1    1    2    1    2    4    2    2    2    1    1    3    3    1    1    1
##  229  230  232  233  234  235  237  238  239  241  242  243  244  247  249  250
##    2    1    1    1    1    1    3    3    1    2    1    2    2    2    3    2
##  253  254  255  257  259  261  262  263  264  266  267  270  275  279  280  282
##    1    2    2    2    1    2    2    1    2    1    1    2    1    2    2    1
##  283  285  286  288  289  291  292  295  296  298  299  301  309  310  315  319
##    2    2    1    2    1    1    1    1    1    2    1    1    1    1    1    1
##  320  331  332  333  334  339  341  344  345  347  348  349  352  354  357  358
##    1    1    4    1    1    1    1    1    1    2    1    1    2    1    1    1
##  363  369  375  376  379  382  386  388  399  404  408  411  414  415  416  419
##    1    1    1    1    1    1    1    1    2    2    1    1    1    1    2    1
##  420  428  433  434  438  439  443  449  453  455  458  460  463  471  482  486
##    1    1    1    2    1    2    1    1    1    1    1    1    1    1    1    1
##  487  488  494  499  503  506  514  515  517  518  520  522  524  525  527  529
##    1    1    1    1    2    1    1    1    1    1    1    1    1    2    1    1
##  531  536  539  541  543  552  561  567  572  578  585  588  589  593  595  599
##    1    1    1    1    1    1    1    1    1    1    2    1    1    1    1    1
##  601  607  622  629  635  645  647  649  661  673  676  687  703  720  731  748
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1108
##    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00   11.00   21.66   24.00 1108.00
##
## includes extended item information - examples:
##                    labels
## 1   *Boombox Ipod Classic
## 2 *USB Office Mirror Ball
## 3                        ?
##
## includes extended transaction information - examples:
##   transactionID
## 1        536365
## 2        536366
## 3        536367
```

```
#For no of transactions
no_trans <- length(order_trans)
```

```r
#For no of unique values
unique_items <- length(itemLabels (order_trans))


cat("No of Transactions:", no_trans, "\n")
```

```
## No of Transactions: 24446
```

```r
cat("Nuo of unique values:", unique_items, "\n")
```

```
## Nuo of unique values: 4211
```

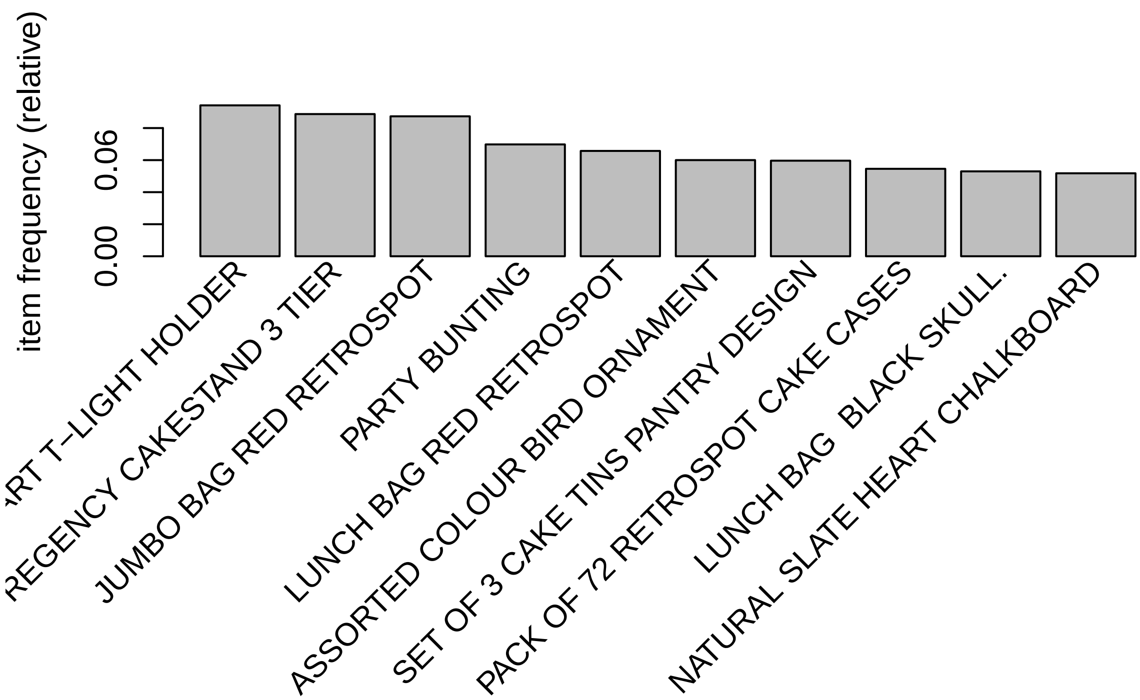
c) Inspect (with inspect) first three transactions. What items are in basket with transaction id 536366? (4 points)

```r
inspect(order_trans[1:3,])
```

```
##     items                                 transactionID
## [1] {CREAM CUPID HEARTS COAT HANGER,
##      GLASS STAR FROSTED T-LIGHT HOLDER,
##      KNITTED UNION FLAG HOT WATER BOTTLE,
##      RED WOOLLY HOTTIE WHITE HEART.,
##      SET 7 BABUSHKA NESTING BOXES,
##      WHITE HANGING HEART T-LIGHT HOLDER,
##      WHITE METAL LANTERN}                        536365
## [2] {HAND WARMER RED POLKA DOT,
##      HAND WARMER UNION JACK}                     536366
## [3] {ASSORTED COLOUR BIRD ORNAMENT,
##      BOX OF 6 ASSORTED COLOUR TEASPOONS,
##      BOX OF VINTAGE ALPHABET BLOCKS,
##      BOX OF VINTAGE JIGSAW BLOCKS,
##      DOORMAT NEW ENGLAND,
##      FELTCRAFT PRINCESS CHARLOTTE DOLL,
##      HOME BUILDING BLOCK WORD,
##      IVORY KNITTED MUG COSY,
##      LOVE BUILDING BLOCK WORD,
##      POPPY'S PLAYHOUSE BEDROOM,
##      POPPY'S PLAYHOUSE KITCHEN,
##      RECIPE BOX WITH METAL HEART}                536367
```

#In the order with transaction id 536366 we have follwoing items HAND WARMER RED POLKA DOT, and HAND WARMER UNION JACK. d) Visualize top 10 frequent items. What is the most frequent? (4 points)

```r
itemFrequencyPlot(order_trans,topN=10)
```



#Most frequent as per plot is WHITE HANGING HEART T-LIGHT HOLDER. e) We want to look at rule which would have at least 100 transactions. What support is corresponding to that? (4 points)

```
temp <- 100/nrow(order_trans)
cat(temp)
```

```
## 0.004090649
```

f) Calculate rules with a rule. Use previously calculated support, confidence of 0.9 and maxlen of 4 (we are looking into the rules with up to 4 items). (4 points)

```
items <- apriori(order_trans,parameter=list(support=0.0041,confidence=0.9,maxlen =4))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.9    0.1    1 none FALSE            TRUE       5  0.0041      1
##  maxlen target  ext
##       4  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
```

```
## Absolute minimum support count: 100
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[4211 item(s), 24446 transaction(s)] done [0.17s].
## sorting and recoding items ... [1558 item(s)] done [0.01s].
## creating transaction tree ... done [0.01s].
## checking subsets of size 1 2 3 4
```

```
## Warning in apriori(order_trans, parameter = list(support = 0.0041, confidence =
## 0.9, : Mining stopped (maxlen reached). Only patterns up to a length of 4
## returned!
```

```
##  done [0.13s].
## writing ... [1216 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```r
items
```

```
## set of 1216 rules
```

#So we have 1216 rules. g) List top 10 by confidence. What is the sense of confidence (explain on example of the top rule)? (4 points)

```r
inspect(head(sort(items,by='confidence'),n=10))
```

```
##      lhs                                         rhs                  support confidence    coverage
## [1]  {CHRISTMAS TREE HEART DECORATION,
##       SUKI  SHOULDER BAG}                     => {DOTCOM POSTAGE} 0.004131555          1 0.004131555 3
## [2]  {PIZZA PLATE IN BOX,
##       SUKI  SHOULDER BAG}                     => {DOTCOM POSTAGE} 0.004417901          1 0.004417901 3
## [3]  {SKULL SHOULDER BAG,
##       URBAN BLACK RIBBONS}                    => {DOTCOM POSTAGE} 0.004172462          1 0.004172462 3
## [4]  {RECYCLING BAG RETROSPOT,
##       SET/4 RED MINI ROSE CANDLE IN BOWL}     => {DOTCOM POSTAGE} 0.004295181          1 0.004295181 3
## [5]  {SET/4 RED MINI ROSE CANDLE IN BOWL,
##       SUKI  SHOULDER BAG}                     => {DOTCOM POSTAGE} 0.004540620          1 0.004540620 3
## [6]  {CHRISTMAS TREE STAR DECORATION,
##       SKULL SHOULDER BAG}                     => {DOTCOM POSTAGE} 0.004295181          1 0.004295181 3
## [7]  {BEADED CRYSTAL HEART GREEN ON STICK,
##       VINTAGE PAISLEY STATIONERY SET}         => {DOTCOM POSTAGE} 0.004540620          1 0.004540620 3
## [8]  {BEADED CRYSTAL HEART GREEN ON STICK,
##       FLORAL FOLK STATIONERY SET}             => {DOTCOM POSTAGE} 0.004867872          1 0.004867872 3
## [9]  {BEADED CRYSTAL HEART GREEN ON STICK,
##       CHARLOTTE BAG SUKI DESIGN}              => {DOTCOM POSTAGE} 0.004131555          1 0.004131555 3
## [10] {BEADED CRYSTAL HEART GREEN ON STICK,
##       SUKI  SHOULDER BAG}                     => {DOTCOM POSTAGE} 0.004745153          1 0.004745153 3
```

From the above data we can say that LHS customer who purchases both {CHRISTMAS TREE HEART DECORATION,SUKI SHOULDER BAG} together also pruchases {DOTCOM POSTAGE} this we can say as the support value is 1. This is the strongest rule in the dataset which we have analysed.

h) List top 10 by lift. What is the sense of lift (explain on example of the top rule)? (4 points)

```r
inspect(head(sort(items,by='lift'),n=10))
```

```
##      lhs                              rhs                           support confidence    coverage
## [1]  {DOLLY GIRL CHILDRENS CUP,
##       SPACEBOY CHILDRENS BOWL,
##       SPACEBOY CHILDRENS CUP}     => {DOLLY GIRL CHILDRENS BOWL} 0.004254275  0.9811321 0.004336088
## [2]  {DOLLY GIRL CHILDRENS CUP,
##       SPACEBOY CHILDRENS BOWL}    => {DOLLY GIRL CHILDRENS BOWL} 0.004826966  0.9593496 0.005031498
## [3]  {DOLLY GIRL CHILDRENS BOWL,
##       SPACEBOY CHILDRENS BOWL,
##       SPACEBOY CHILDRENS CUP}     => {DOLLY GIRL CHILDRENS CUP}  0.004254275  0.9541284 0.004458807
## [4]  {DOLLY GIRL CHILDRENS BOWL,
##       SPACEBOY CHILDRENS CUP}     => {DOLLY GIRL CHILDRENS CUP}  0.004581527  0.9411765 0.004867872
## [5]  {DOLLY GIRL CHILDRENS BOWL,
##       DOLLY GIRL CHILDRENS CUP,
##       SPACEBOY CHILDRENS CUP}     => {SPACEBOY CHILDRENS BOWL}   0.004254275  0.9285714 0.004581527
## [6]  {DOLLY GIRL CHILDRENS BOWL,
##       SPACEBOY CHILDRENS CUP}     => {SPACEBOY CHILDRENS BOWL}   0.004458807  0.9159664 0.004867872
## [7]  {HERB MARKER BASIL,
##       HERB MARKER CHIVES,
##       HERB MARKER ROSEMARY}       => {HERB MARKER THYME}         0.006913196  0.9825581 0.007035916
## [8]  {PINK VINTAGE SPOT BEAKER,
##       RED VINTAGE SPOT BEAKER}    => {BLUE VINTAGE SPOT BEAKER}  0.004581527  0.9256198 0.004949685
## [9]  {HERB MARKER CHIVES,
##       HERB MARKER MINT,
##       HERB MARKER ROSEMARY}       => {HERB MARKER PARSLEY}       0.007158635  0.9831461 0.007281355
## [10] {HERB MARKER CHIVES,
##       HERB MARKER MINT,
##       HERB MARKER THYME}          => {HERB MARKER PARSLEY}       0.007117729  0.9830508 0.007240448
```

#In the first row we can see lift value as 122.98 ~ 123 which means that a very strong positive association between the items on LHS and ITEMS AND rhs. This indicates that the probability of purchasing a "DOLLY GIRL CHILDRENS BOWL" in same transACSTIN is increased by 123 times when "SPACEBOY CHILDRENS BOWL" and "DOLLY GIRL CHILDRENS CUP" are there in the transaction, as opposed to when the purchase of the "DOLLY GIRL CHILDRENS BOWL" is made separately from the first two items.

LHS = {DOLLY GIRL CHILDRENS CUP,
SPACEBOY CHILDRENS BOWL,
SPACEBOY CHILDRENS CUP}

RHS= {DOLLY GIRL CHILDRENS BOWL}