EAS509 Exam 2

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.

```
library(tibble)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(readr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
library(tsibble)
```

```
## Warning: package 'tsibble' was built under R version 4.2.2
```

```
##
## Attaching package: 'tsibble'
```

```
## The following object is masked from 'package:lubridate':
##
##     interval
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, union
```

```
library(fable)
```

```
## Warning: package 'fable' was built under R version 4.2.2
```

```
## Loading required package: fabletools
```

```
## Warning: package 'fabletools' was built under R version 4.2.2
```

```
library(fabletools)
library(feasts)
```

```
## Warning: package 'feasts' was built under R version 4.2.2
```

```
library(tsibbledata)
```

```
## Warning: package 'tsibbledata' was built under R version 4.2.2
```

```
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.2.2
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
library(arules)
```

```
## Warning: package 'arules' was built under R version 4.2.2
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Attaching package: 'arules'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(arulesViz)
```

```
## Warning: package 'arulesViz' was built under R version 4.2.2
```

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:** 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:** 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:** 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:** 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:** all should present for full score Residuals have mean zero, Residuals are normally distributed, Residuals have constant variance

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:** 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:** Forecasts are biased

8. Which measure of forecast accuracy is scale independent?

a. MAE
b. MSE
c. RMSE
d. MAPE **Answer:** MAPE

9. Calculation of forecasts is based on what?

a. Test set
b. Training set
c. Both
d. Neither **Answer:** Training set

10. Forecast accuracy is based on what?

a. Test set
b. Training set
c. Both
d. Neither **Answer:** 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:** 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:** either or all is ok for full credit Cyclical

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

a. Trend
b. Seasonal
c. Cyclical
d. White Noise **Answer:** White Noise

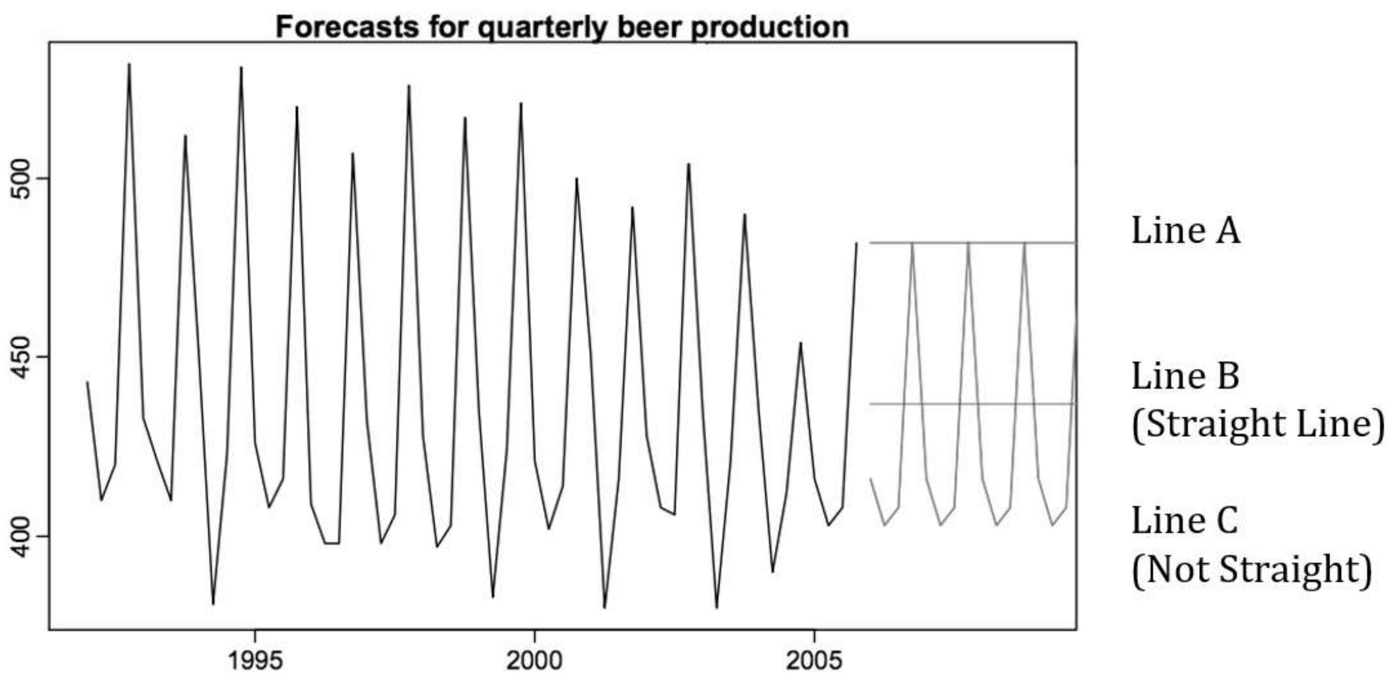
14. Which of the following time series decomposition models is appropriate when the magnitude of the seasonal fluctuations are not proportional to the level?

a. Additive
b. Multiplicative
c. Both
d. Neither **Answer:** 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:** Multiplicative

Exhibit 1



16. Refer to Exhibit 1. Line A is which simple forecasting method? a. Average Method b. Naïve Method c. Seasonal Naïve Method d. Drift **Answer:** Naive Method

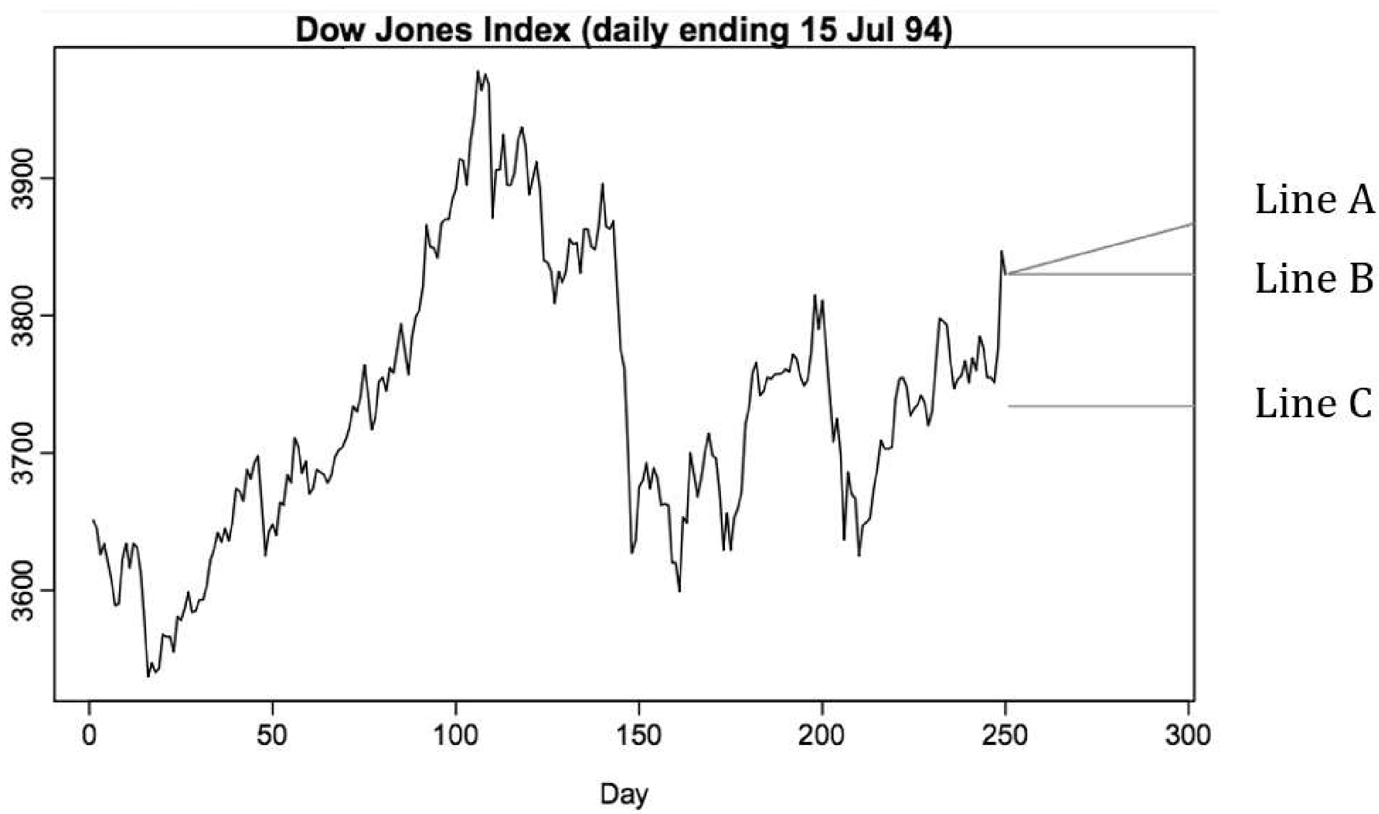
17. Refer to Exhibit 1. Line B is which simple forecasting method?

a. Average Method
b. Naïve Method
c. Seasonal Naïve Method
d. Drift Method **Answer:** Average Method

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

a. Average Method
b. Naïve Method
c. Seasonal Naïve Method
d. Drift Method **Answer:** Seasonal Naive 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:** Drift Method

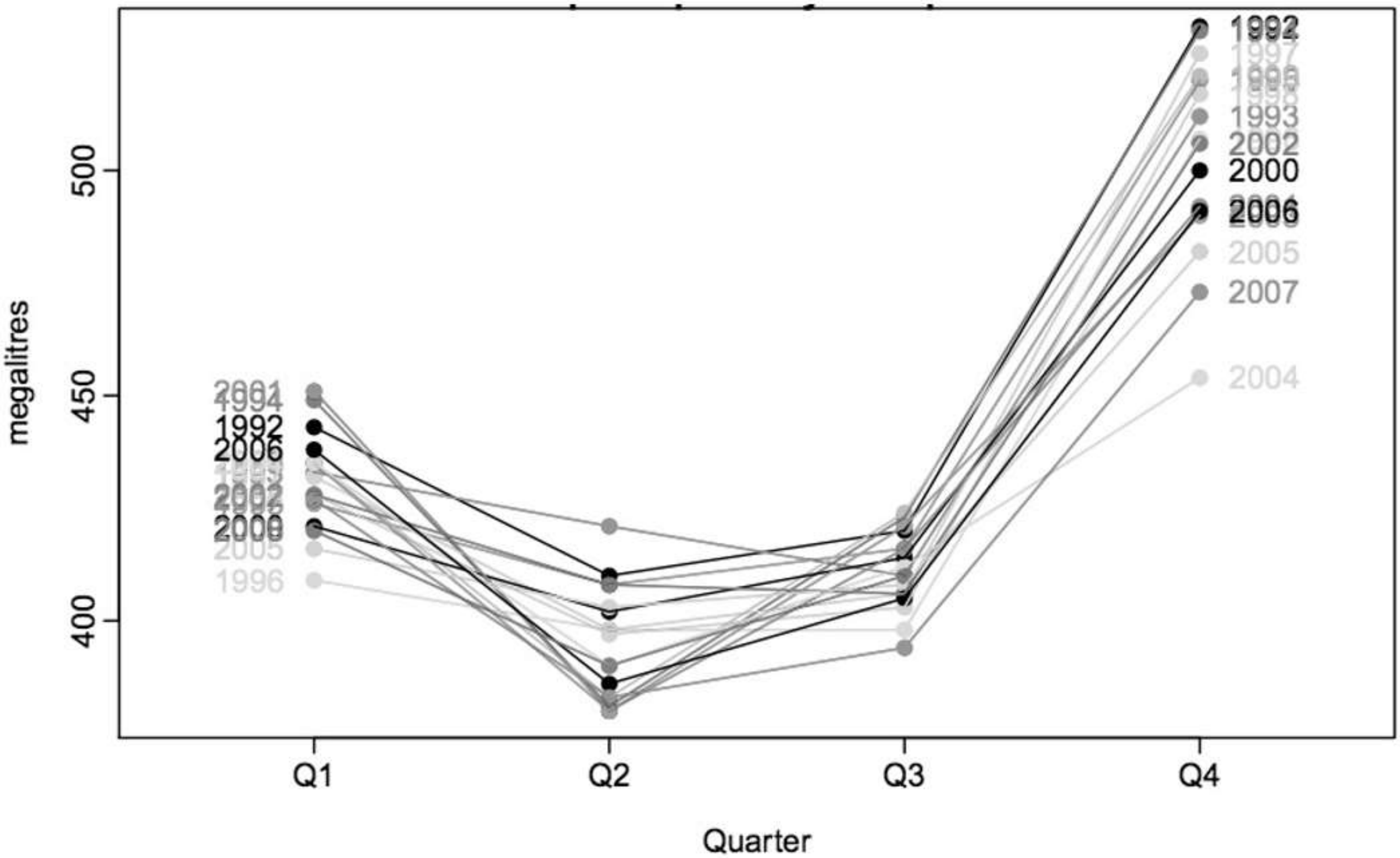
20. Refer to Exhibit 2. Line B is which simple forecasting method?

a. Average Method
b. Naïve Method
c. Seasonal Naïve Method
d. Drift Method **Answer:** Naive 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:** Average Method

Exhibit 3

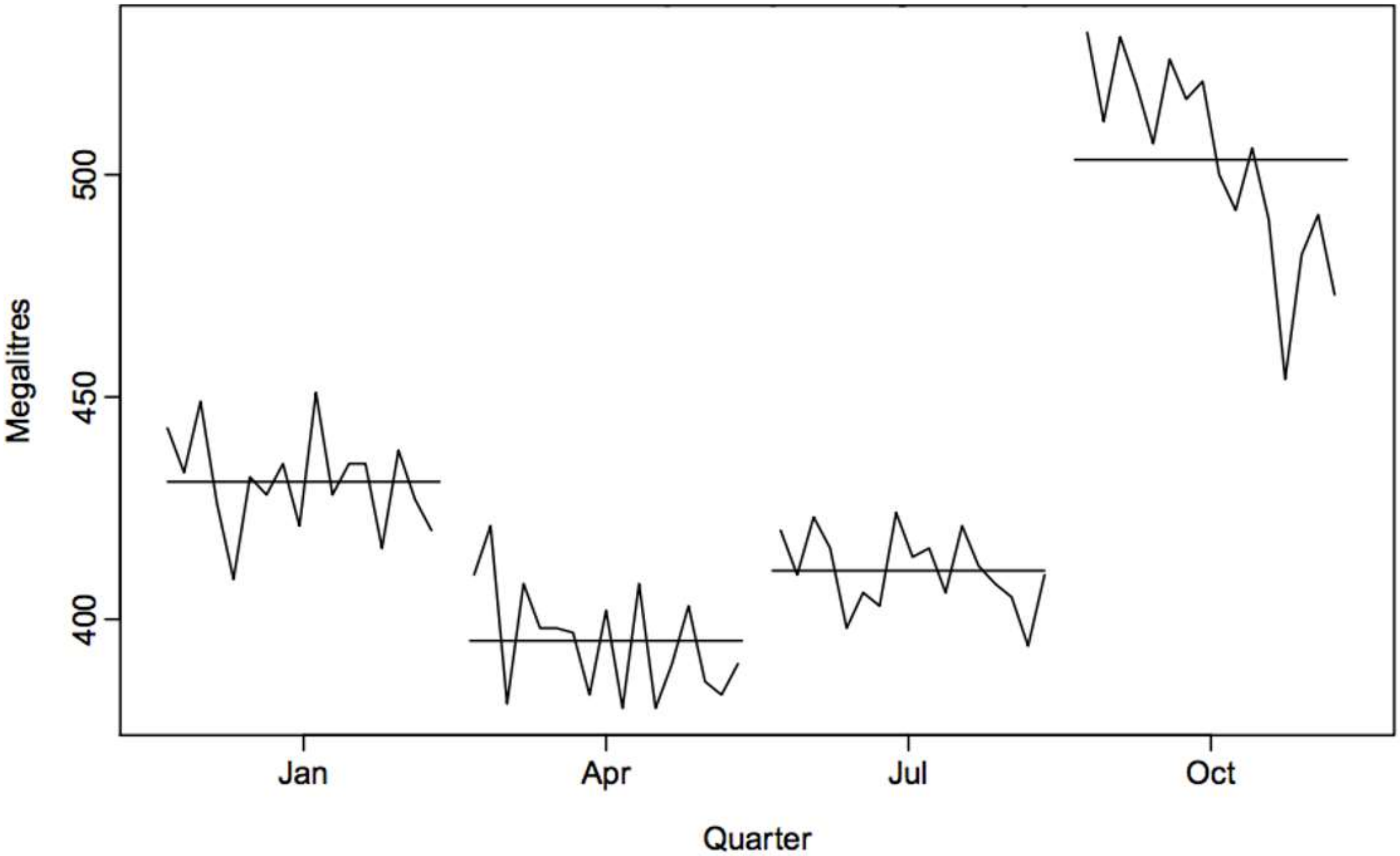


22. Refer to Exhibit 3. The peaks are in which quarter? a. Quarter 1 b. Quarter 2 c. Quarter 3 d. Quarter 4
Answer: Quarter 4

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

a. Quarter 1
b. Quarter 2
c. Quarter 3
d. Quarter 4 **Answer:** there are few in Q3 but largly it is Q2 Quarter 2

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: 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:** 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:** 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:** 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:** 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 **Answer:** 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:** 13.85

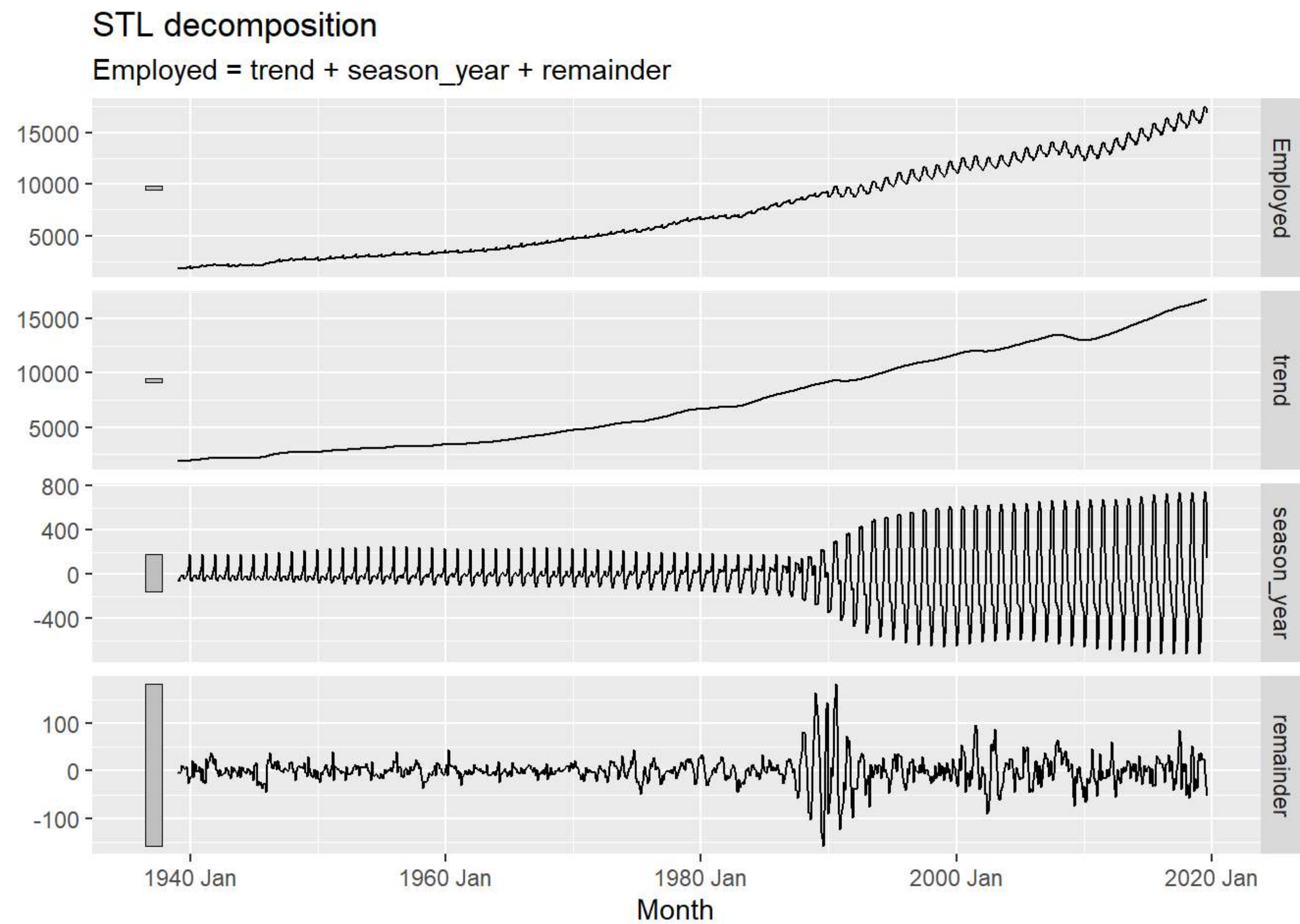
Part B (30 points)

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

```
df = read.csv("us_employment.cvs")
df <- df %>% filter(Title == 'Leisure and Hospitality')
df %>% mutate(Month=yearmonth(Month)) %>% tsibble(index=Month) -> df
df_emp <- select(df,Month,Employed)
```

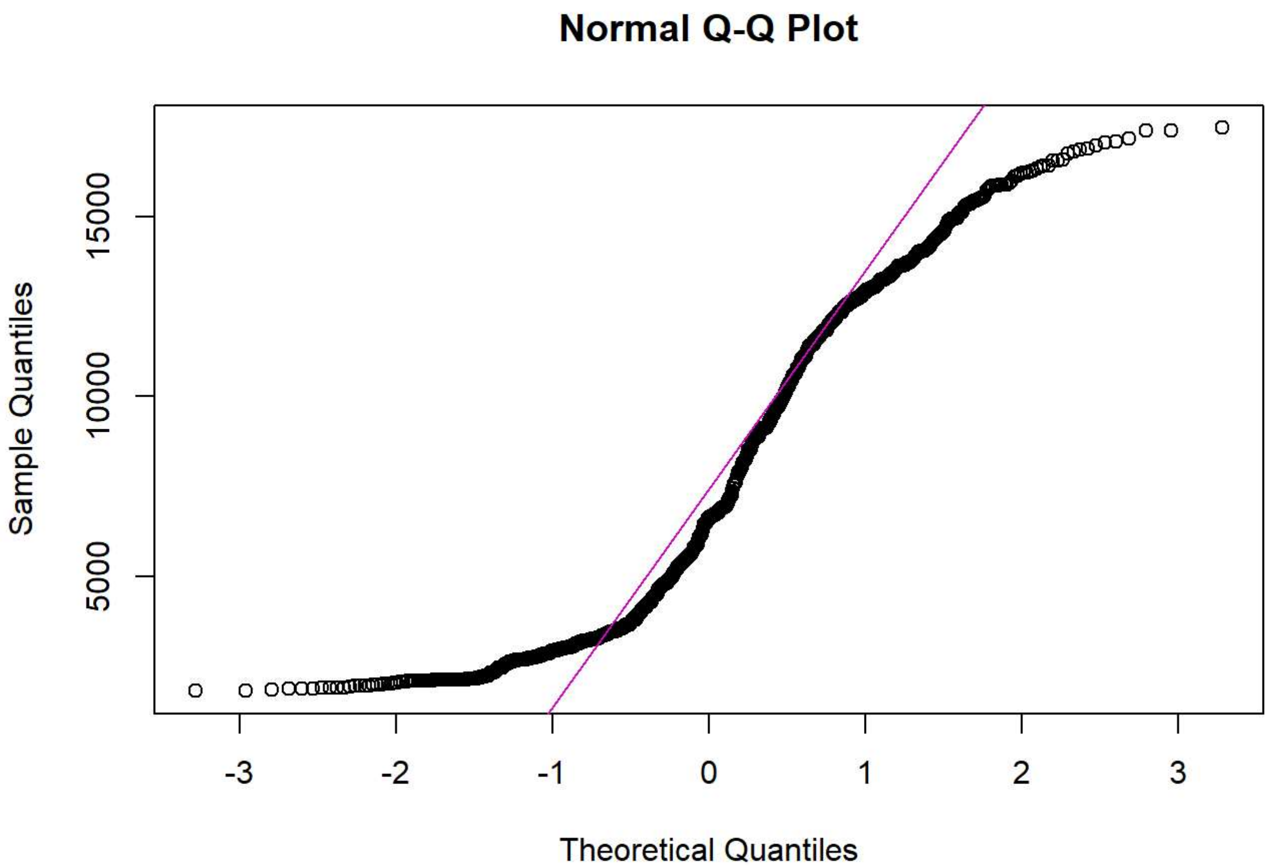
a. Produce an STL decomposition of the data and describe the trend and seasonality. (4 points)

```
df_emp %>% model(STL(Employed)) %>% components() %>% autoplot()
```



b. Do the data need transforming? If so, find a suitable transformation.(4 points)

```
qqnorm(df_emp$Employed)
qqline(df_emp$Employed, col = 14)
```

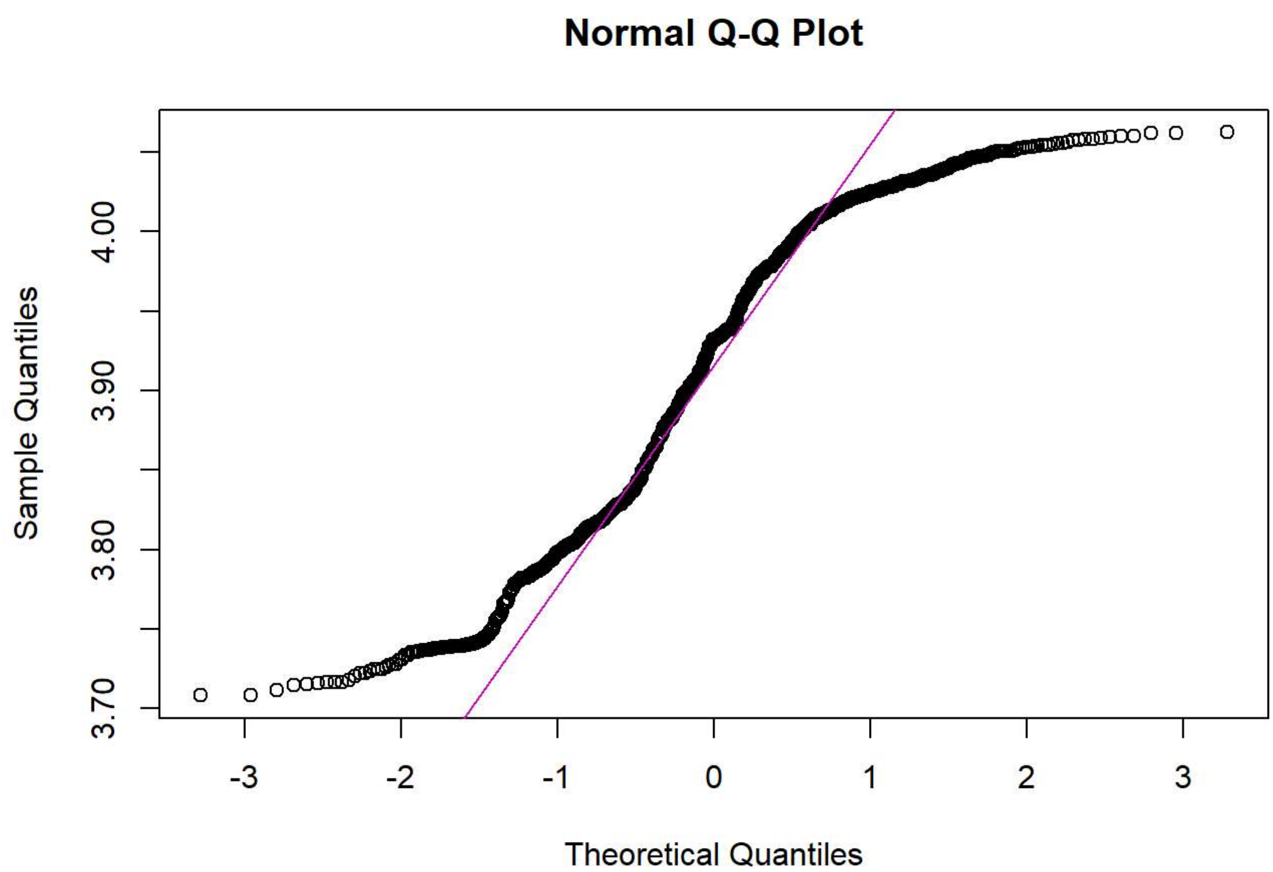


Yes,

the data can be transformed, we will perform Box transform

```
# BoxCox transformation
trans_emp_df <- df_emp %>% as_tsibble() %>% mutate(Employed = box_cox(Employed,df_emp %>% featur
es(Employed, features = guerrero)))

# Checking with the qq plot
qqnorm(trans_emp_df$Employed)
qqline(trans_emp_df$Employed, col = 14)
```



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

```
# KPSS test
trans_emp_df %>%
features(Employed, unitroot_kpss)
```

kpss_stat <dbl>	kpss_pvalue <dbl>
12.1348	0.01

1 row

```
# ndiffs to check the order for differencing
trans_emp_df %>% features(Employed, unitroot_nsdiffs)
```

nsdiffs <int>
1

1 row

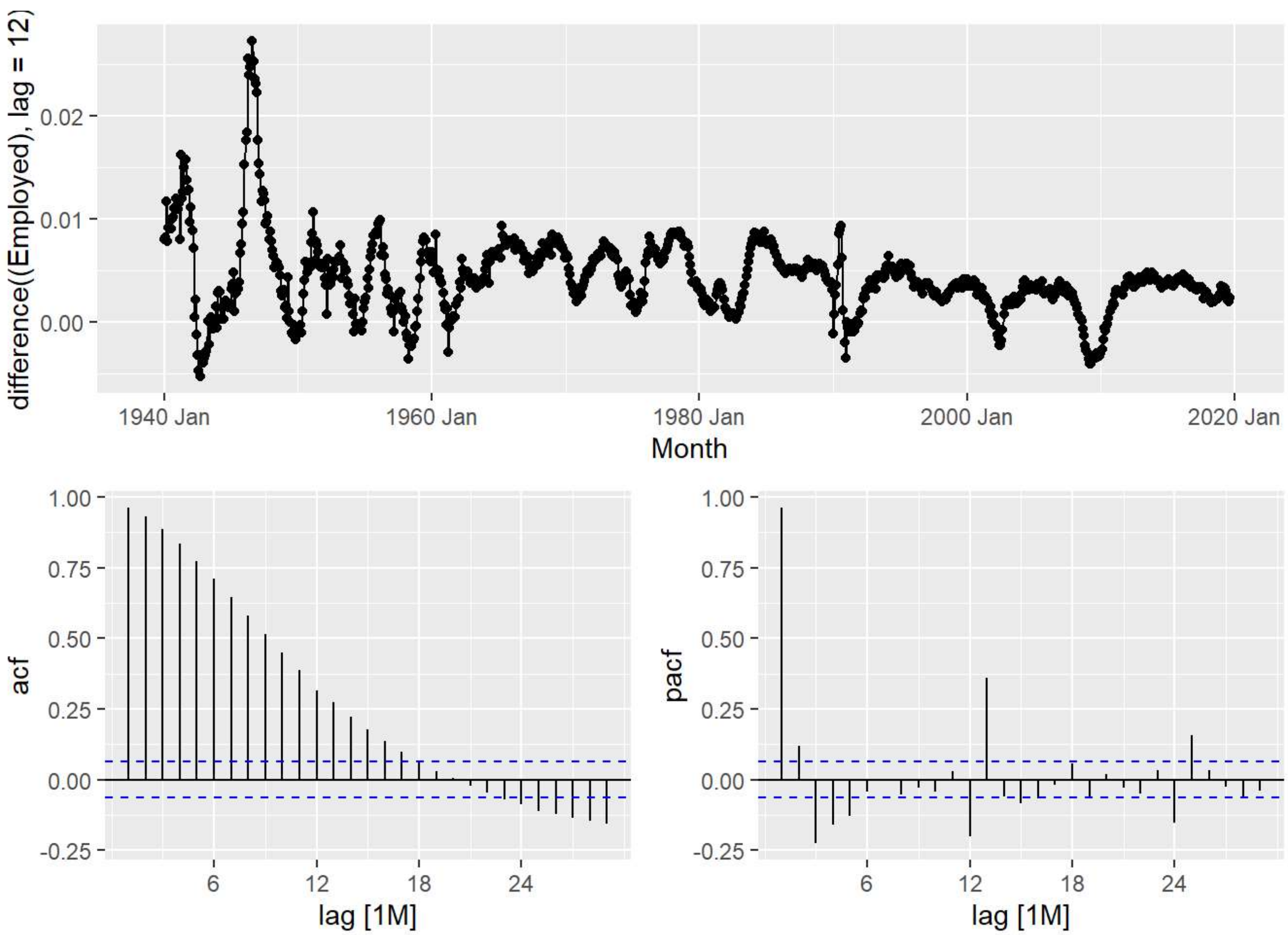
The p-value of KPSS test is less than 0.05, hence data is not stationary. The order for differencing using ndiffs is 1.

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)

```
gg_tsdisplay(trans_emp_df, difference((Employed), lag=12), plot_type='partial')
```

```
## Warning: Removed 12 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```




```
set.seed(47)
# Fit the ARIMA models
fit_emp <- trans_emp_df %>%
  model(
    arima_auto_emp = ARIMA(log(Employed), stepwise = FALSE, approx = FALSE),
    arima1_emp = ARIMA(log(Employed)~0 + pdq(1,1,2) + PDQ(1,1,1))
  )
accuracy(fit_emp)
```

.model	.type	ME	RMSE	MAE	MPE	MA
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<(
arima_auto_emp	Training	-1.933441e-05	0.001045638	0.0005815295	-0.0005083962	0.01505

.model	.type	ME	RMSE	MAE	MPE	MA
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<(
arima1_emp	Training	-2.050250e-05	0.001060544	0.0005866131	-0.0005390780	0.01519

2 rows | 1-8 of 10 columns

```
report(fit_emp[1])
```

```
## Series: Employed
## Model: ARIMA(1,1,2)(2,1,1)[12]
## Transformation: log(Employed)
##
## Coefficients:
##          ar1      ma1     ma2      sar1     sar2    sma1
##       0.6018  -0.6506  0.2543  -1.0882  -0.4728  0.6035
## s.e.  0.0751   0.0742  0.0345   0.0664   0.0349  0.0697
##
## sigma^2 estimated as 7.717e-08:  log likelihood=6646.06
## AIC=-13278.11   AICc=-13277.99   BIC=-13244.07
```

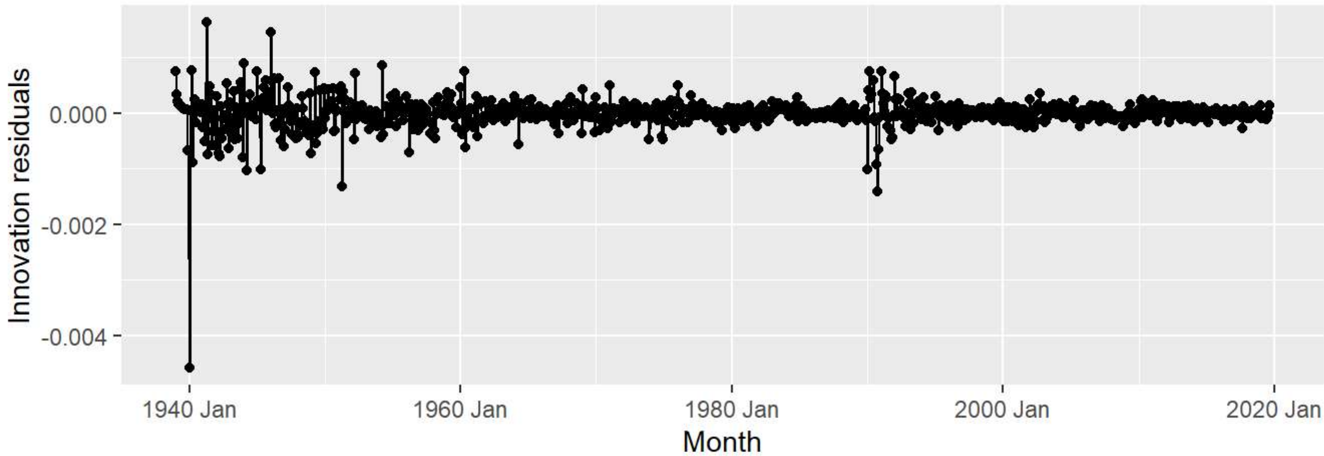
```
report(fit_emp[2])
```

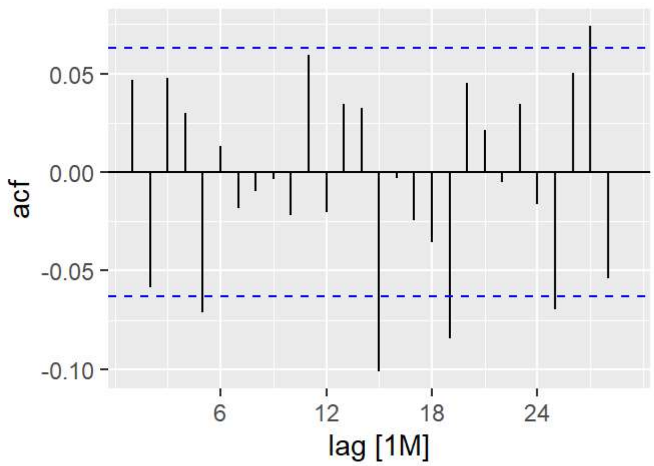
```
## Series: Employed
## Model: ARIMA(1,1,2)(1,1,1)[12]
## Transformation: log(Employed)
##
## Coefficients:
##          ar1      ma1     ma2      sar1     sma1
##       0.5685  -0.6452  0.2692  -0.1501  -0.3550
## s.e.  0.0768   0.0756  0.0332   0.0705   0.0665
##
## sigma^2 estimated as 7.933e-08:  log likelihood=6627.43
## AIC=-13242.86   AICc=-13242.77   BIC=-13213.68
```

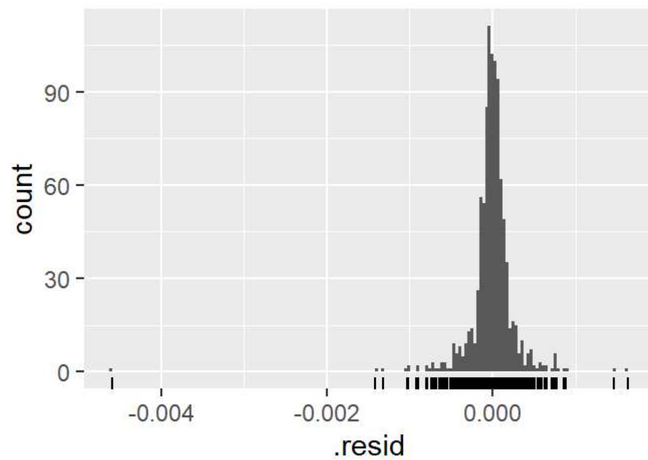
AICc of the auto arima model is -13278.11, it is lower than the manualli created Arima model. So, the auto arima model having (1,1,2)(2,1,1)[12] is the best.

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)

```
gg_tsresiduals(fit_emp %>% select(arima_auto_emp))
```








```
augment(fit_emp) %>%  features(.innov, ljung_box, lag=24, dof=7)
```

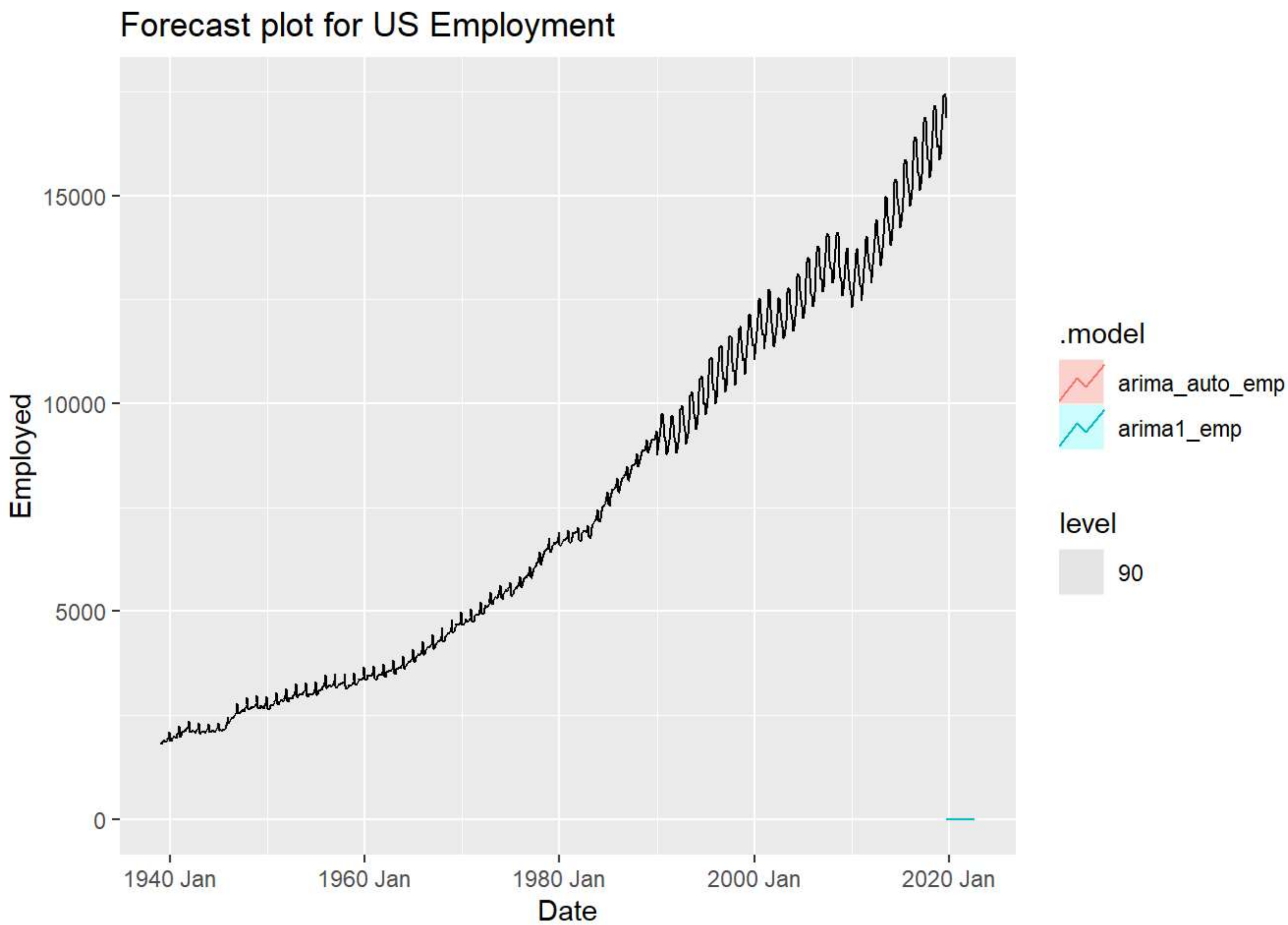
.model <chr>	lb_stat <dbl>	lb_pvalue <dbl>
arima_auto_emp	43.62548	0.0003888467
arima1_emp	43.40074	0.0004195168

2 rows

Using the ACF plot we can tell that most lag lies within the boundary, only three of them crosss it. Hence, there is some white noise. Our p value is also close to 0.0003888467 so our model has white noise.

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

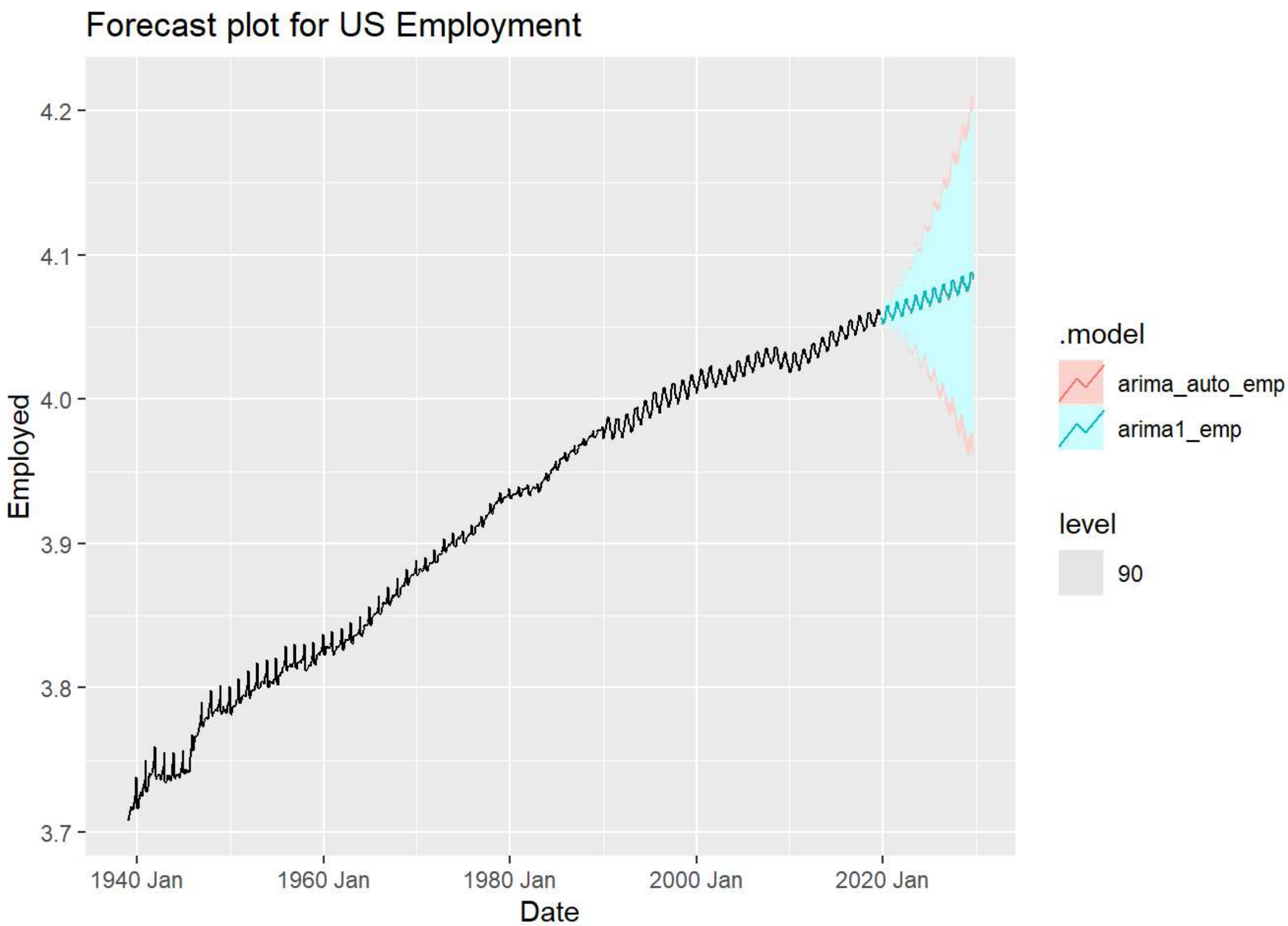
```
arima_auto_emp_fc <- fit_emp %>% forecast(h = "3 years")
arima_auto_emp_fc %>% autoplot(df_emp,level = 90) +labs(title="Forecast plot for US Employment",
x = 'Date', y="Employed")
```



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)

```
arima_auto_emp_fc <- fit_emp %>% forecast(h = "10 years")
arima_auto_emp_fc %>% autoplot(trans_emp_df,level = 90) +labs(title="Forecast plot for US Employ
ment", x = 'Date', y="Employed")
```



We are able to approximately forecast the data for 5 years correctly.After that it is not reliable.

Part C (40 points)

Consider following transactions: (8 points)

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  = N_bananas/N= 0.4
confidence = N_bananas_yogurt/N_bananas = 1

support_yogurt =N_yogurt/N = 0.6

lift = confidence/support_yogurt = 1.67
```

• {Corn, Bread}->{Onion} (3 points)

```
N= 5
N_corn = 2
N_bread = 5
N_onion = 3
N_bread_onion = 3
N_corn_bread_onion = 1

support  = N_corn_bread_onion/N = 0.2
confidence = N_corn_bread_onion/N_bread_onion = 0.33

support_onion = N_onion/N = 0.6

lift = confidence/support_onion = 0.55
```

• {Bread}->{Milk, Yogurt} (3 points)

```
N= 5
N_bread = 5
N_milk = 3
N_yogurt = 3
N_milk_yogurt = 2
N_bread_milk_yougurt = 2

support  = N_bread_milk_yougurt/N = 0.4
confidence = N_bread_milk_yougurt/N_bread = 0.4

support_milk_yogurt = N_milk_yogurt/N = 0.4

lift = confidence/support_milk_yogurt = 1
```

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)

```
df = read.csv("Online_Retail2.csv")
df = df %>% select(-c(StockCode, InvoiceDate, CustomerID, Country))
trans_df <- transactions(df, format='long')
```

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

```
summary(trans_df)
```

```
## 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
```

TOtal = 24446 transactions with unique items = 4211. c) Inspect (with inspect) first three transactions. What items are in basket with transaction id 536366? (4 points)

```
inspect(head(trans_df, n = 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
```

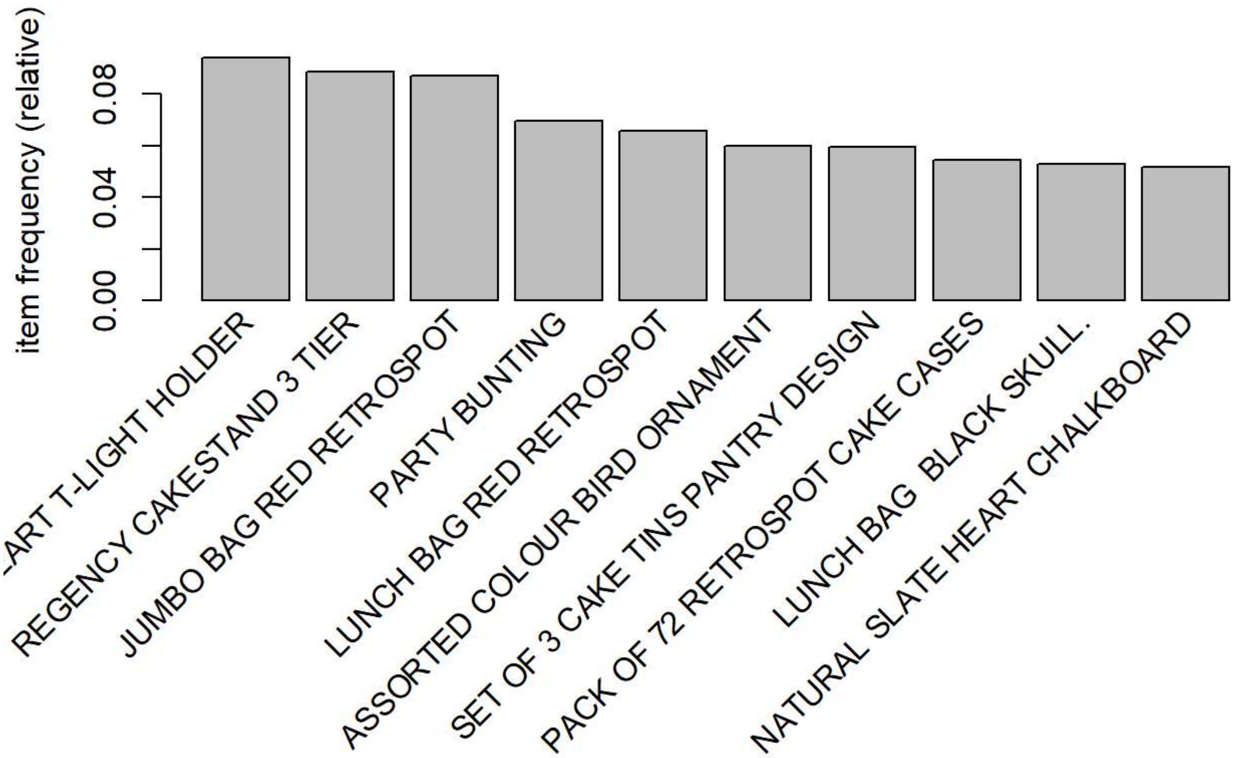
```
df[df$InvoiceNo=="536366",]
```

	InvoiceNo <chr>	Description <chr>	Quantity <int>	UnitPrice <dbl>
8	536366	HAND WARMER UNION JACK	6	1.85
9	536366	HAND WARMER RED POLKA DOT	6	1.85

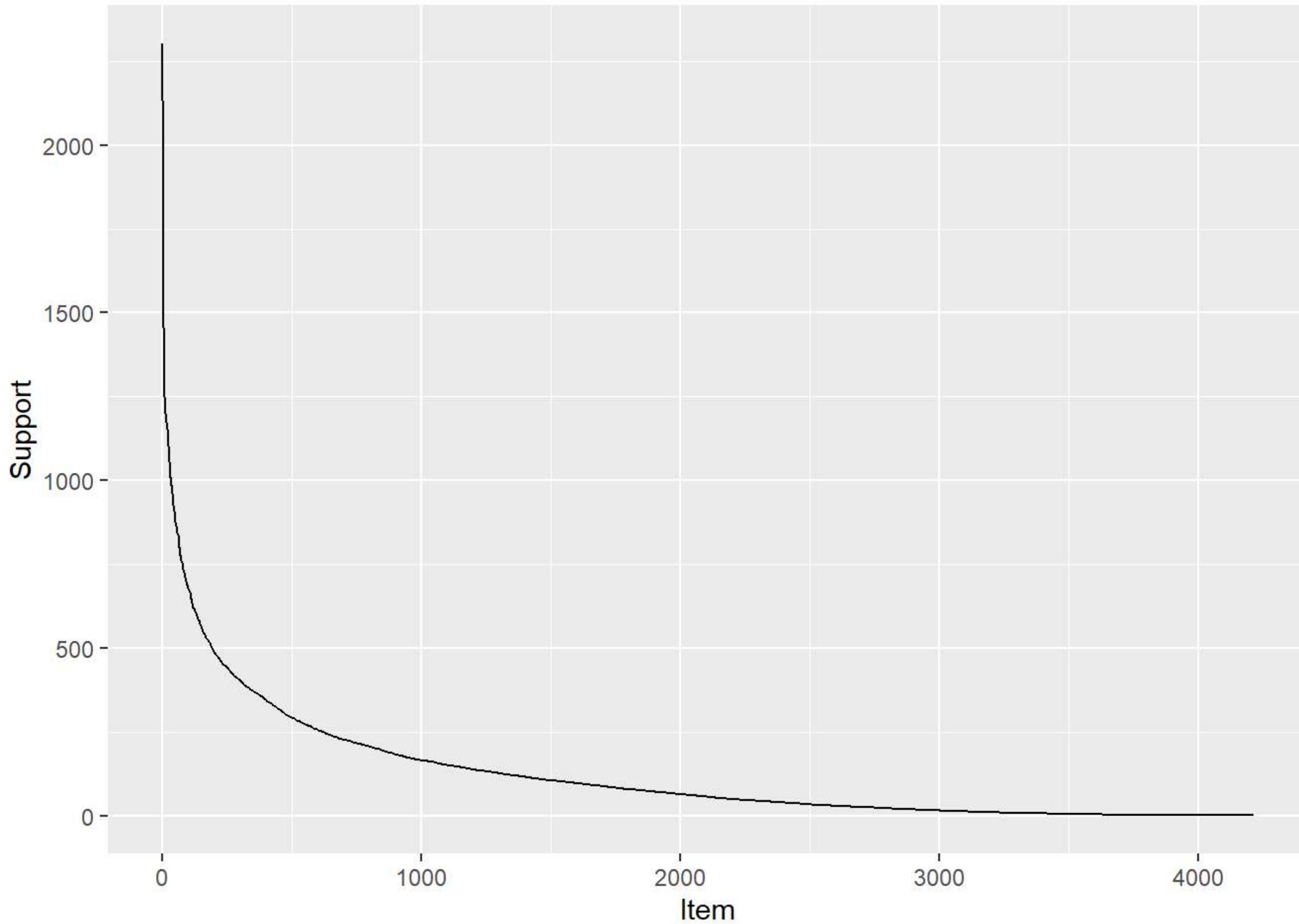
2 rows

HAND WARMER RED POLKA DOT and HAND WARMER UNION JACK. d) Visualize top 10 frequent items. What is the most frequent? (4 points)

```
itemFrequencyPlot(trans_df,topN = 10)
```




```
ggplot(tibble(Support = sort(itemFrequency(trans_df, type = "absolute"), decreasing = TRUE),
    Item = seq_len(ncol(trans_df))
  ), aes(x = Item, y = Support)) + geom_line()
```



HANGING HEART T-LIGHT HOLDER is the most frequent e) We want to look at rule which would have at least 100 transactions. What support is corresponding to that? (4 points)

```
100/nrow(trans_df)
```

```
## [1] 0.004090649
```

```
rule_retail = apriori(trans_df, parameter = list(support = 0.004090649, confidence = 0.9))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime     support minlen
##         0.9    0.1    1 none FALSE            TRUE       5 0.004090649      1
##  maxlen target  ext
##      10  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 [0.13s].
## sorting and recoding items ... [1558 item(s)] done [0.01s].
## creating transaction tree ... done [0.01s].
## checking subsets of size 1 2 3 4 5 6 7 done [0.13s].
## writing ... [1603 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
inspect(head(rule_retail, n=1))
```

```
##     lhs                       rhs                 support     confidence
## [1] {HERB MARKER CHIVES} => {HERB MARKER MINT} 0.007813139 0.9095238
##     coverage    lift     count
## [1] 0.008590362 92.25817 191
```

ANSWER 0.00409 is the support value for 100 transaction .

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)

```
cal_rules <- apriori(trans_df, parameter = list(supp=0.004090649, conf=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 0.004090649      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 [0.15s].
## 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(trans_df, parameter = list(supp = 0.004090649, conf = 0.9, :
## Mining stopped (maxlen reached). Only patterns up to a length of 4 returned!
```

```
##  done [0.12s].
## writing ... [1216 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
summary(cal_rules)
```

```
## set of 1216 rules
##
## rule length distribution (lhs + rhs):sizes
##   2   3   4
##  12 483 721
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.000   4.000   3.583   4.000   4.000
##
## summary of quality measures:
##     support           confidence        coverage             lift
##  Min.   :0.004132   Min.   :0.9000   Min.   :0.004132   Min.   : 10.33
##  1st Qu.:0.004295   1st Qu.:0.9182   1st Qu.:0.004541   1st Qu.: 31.58
##  Median :0.004582   Median :0.9430   Median :0.004868   Median : 33.31
##  Mean   :0.005023   Mean   :0.9471   Mean   :0.005316   Mean   : 38.96
##  3rd Qu.:0.005277   3rd Qu.:0.9766   3rd Qu.:0.005604   3rd Qu.: 34.48
##  Max.   :0.010554   Max.   :1.0000   Max.   :0.011536   Max.   :123.00
##      count
##  Min.   :101.0
##  1st Qu.:105.0
##  Median :112.0
##  Mean   :122.8
##  3rd Qu.:129.0
##  Max.   :258.0
##
## mining info:
##      data ntransactions     support confidence
##  trans_df         24446 0.004090649        0.9
##                                                                                       call
##  apriori(data = trans_df, parameter = list(supp = 0.004090649, conf = 0.9, maxlen = 4))
```

Changing maxlen to 4 we get a total of 1216 rules.

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

```
inspect(head(cal_rules, n = 10, by = "confidence"))
```

```
##      lhs                                             rhs                    support confidence    cov
erage     lift count
## [1]  {CHRISTMAS TREE HEART DECORATION,
##       SUKI  SHOULDER BAG}                       => {DOTCOM POSTAGE} 0.004131555          1 0.0041
31555 34.47955   101
## [2]  {PIZZA PLATE IN BOX,
##       SUKI  SHOULDER BAG}                       => {DOTCOM POSTAGE} 0.004417901          1 0.0044
17901 34.47955   108
## [3]  {SKULL SHOULDER BAG,
##       URBAN BLACK RIBBONS}                      => {DOTCOM POSTAGE} 0.004172462          1 0.0041
72462 34.47955   102
## [4]  {RECYCLING BAG RETROSPOT,
##       SET/4 RED MINI ROSE CANDLE IN BOWL}   => {DOTCOM POSTAGE} 0.004295181          1 0.0042
95181 34.47955   105
## [5]  {SET/4 RED MINI ROSE CANDLE IN BOWL,
##       SUKI  SHOULDER BAG}                       => {DOTCOM POSTAGE} 0.004540620          1 0.0045
40620 34.47955   111
## [6]  {CHRISTMAS TREE STAR DECORATION,
##       SKULL SHOULDER BAG}                       => {DOTCOM POSTAGE} 0.004295181          1 0.0042
95181 34.47955   105
## [7]  {BEADED CRYSTAL HEART GREEN ON STICK,
##       VINTAGE PAISLEY STATIONERY SET}       => {DOTCOM POSTAGE} 0.004540620          1 0.0045
40620 34.47955   111
## [8]  {BEADED CRYSTAL HEART GREEN ON STICK,
##       FLORAL FOLK STATIONERY SET}           => {DOTCOM POSTAGE} 0.004867872          1 0.0048
67872 34.47955   119
## [9]  {BEADED CRYSTAL HEART GREEN ON STICK,
##       CHARLOTTE BAG SUKI DESIGN}            => {DOTCOM POSTAGE} 0.004131555          1 0.0041
31555 34.47955   101
## [10] {BEADED CRYSTAL HEART GREEN ON STICK,
##       SUKI  SHOULDER BAG}                       => {DOTCOM POSTAGE} 0.004745153          1 0.0047
45153 34.47955   116
```

CHRISTMAS TREE HEART DECORATION, SUKI SHOULDER BAG} => {DOTCOM POSTAGE} is our the rule with confidence 1. From this we can tell that DOTCOM POSTAGE appear in transactions that contain CHRISTMAS TREE HEART DECORATION and SUKI SHOULDER BAG quite often. h) List top 10 by lift. What is the sense of lift (explain on example of the top rule)? (4 points)

```
inspect(head(cal_rules, n = 10, by = "lift"))
```

```
##      lhs                                rhs                              support confidence    co
verage     lift count
## [1]  {DOLLY GIRL CHILDRENS CUP,                                                                 
##       SPACEBOY CHILDRENS BOWL,                                                                  
##       SPACEBOY CHILDRENS CUP}     => {DOLLY GIRL CHILDRENS BOWL} 0.004254275  0.9811321 0.004
336088 122.9987   104
## [2]  {DOLLY GIRL CHILDRENS CUP,                                                                 
##       SPACEBOY CHILDRENS BOWL}    => {DOLLY GIRL CHILDRENS BOWL} 0.004826966  0.9593496 0.005
031498 120.2680   118
## [3]  {DOLLY GIRL CHILDRENS BOWL,                                                                
##       SPACEBOY CHILDRENS BOWL,                                                                  
##       SPACEBOY CHILDRENS CUP}     => {DOLLY GIRL CHILDRENS CUP}  0.004254275  0.9541284 0.004
458807 117.8011   104
## [4]  {DOLLY GIRL CHILDRENS BOWL,                                                                
##       SPACEBOY CHILDRENS CUP}     => {DOLLY GIRL CHILDRENS CUP}  0.004581527  0.9411765 0.004
867872 116.2020   112
## [5]  {DOLLY GIRL CHILDRENS BOWL,                                                                
##       DOLLY GIRL CHILDRENS CUP,                                                                 
##       SPACEBOY CHILDRENS CUP}     => {SPACEBOY CHILDRENS BOWL}   0.004254275  0.9285714 0.004
581527 105.0919   104
## [6]  {DOLLY GIRL CHILDRENS BOWL,                                                                
##       SPACEBOY CHILDRENS CUP}     => {SPACEBOY CHILDRENS BOWL}   0.004458807  0.9159664 0.004
867872 103.6653   109
## [7]  {HERB MARKER BASIL,                                                                        
##       HERB MARKER CHIVES,                                                                       
##       HERB MARKER ROSEMARY}       => {HERB MARKER THYME}         0.006913196  0.9825581 0.007
035916 100.9228   169
## [8]  {PINK VINTAGE SPOT BEAKER,                                                                 
##       RED VINTAGE SPOT BEAKER}    => {BLUE VINTAGE SPOT BEAKER}  0.004581527  0.9256198 0.004
949685 100.5676   112
## [9]  {HERB MARKER CHIVES,                                                                       
##       HERB MARKER MINT,                                                                         
##       HERB MARKER ROSEMARY}       => {HERB MARKER PARSLEY}       0.007158635  0.9831461 0.007
281355 100.5606   175
## [10] {HERB MARKER CHIVES,                                                                       
##       HERB MARKER MINT,                                                                         
##       HERB MARKER THYME}          => {HERB MARKER PARSLEY}       0.007117729  0.9830508 0.007
240448 100.5509   174
```

{DOLLY GIRL CHILDRENS CUP,
SPACEBOY CHILDRENS BOWL,
SPACEBOY CHILDRENS CUP} => {DOLLY GIRL CHILDRENS BOWL} is the top rule. Lift is the measure of probability of how DOLLY GIRL CHILDRENS BOWL changes if DOLLY GIRL CHILDRENS CUP, SPACEBOY CHILDRENS BOWL and SPACEBOY CHILDRENS CUP are observed.