Data 624: Week 1 Homework

Angrand, Burke, Deboch, Groysman, Karr

October 12, 2019

Week 1 Assignment

HW - Chapter 2 HA 2.1, 2.3

*2.1 Use the help function to explore what the series **gold**, **woolyrnq** and **gas** represent.*

Evaluation of *gold*:

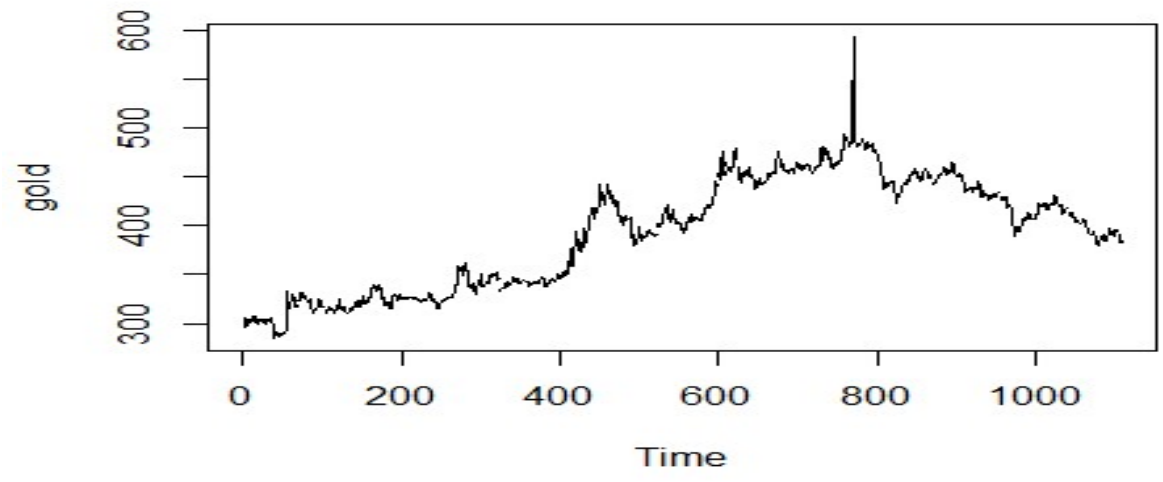
```
help(gold)
```

```
## starting httpd help server ... done
```

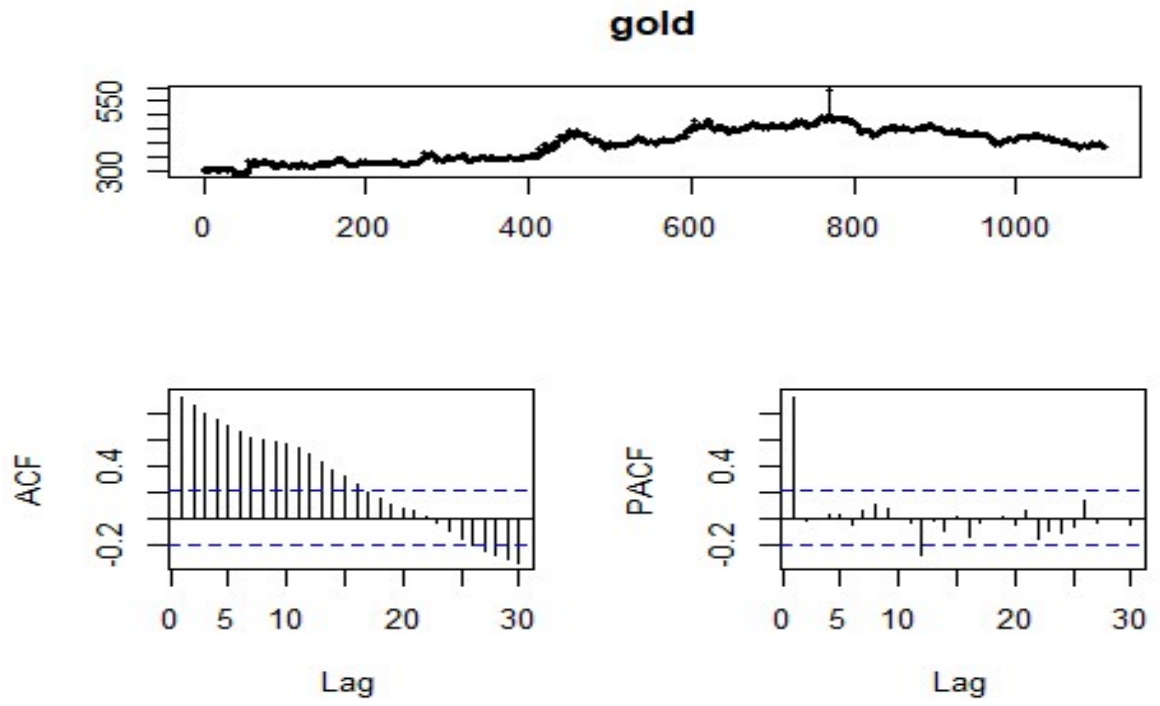
```
#describe(gold)
head(gold)
```

```
## Time Series:
## Start = 1
## End = 6
## Frequency = 1
## [1] 306.25 299.50 303.45 296.75 304.40 298.35
```

```
plot(gold)
```



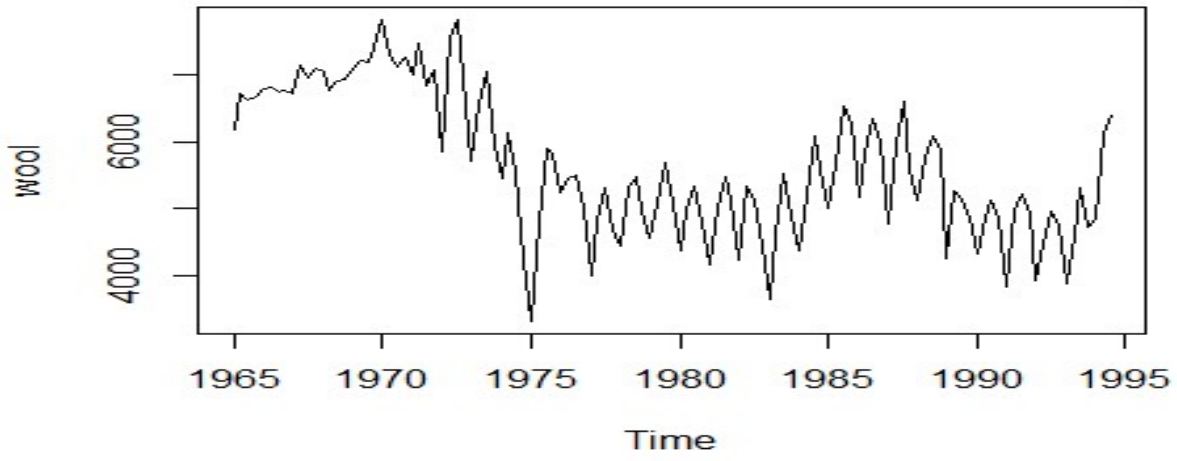

```
tsdisplay(gold)
```



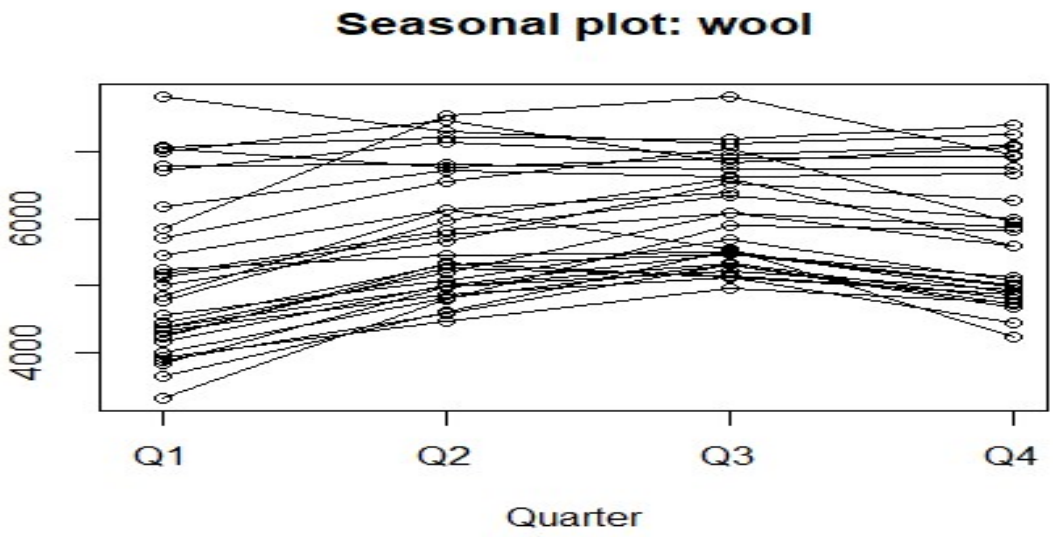
- The gold dataset is structured as a timeseries object. The description notes that the dimensions being compared are time in days 1 January 1985 - 31 March 1989 against price in gold of US dollar. Based on the plot, the price of gold steadily increases until ~800 days pass, there is a notable 30% spike and then a steady dropoff in price for the remaining 250 days.
- The time dimension of this dataset being daily, isn't setup to identify seasonality or year-over-year trends, so this type of evaluation has not been done. It is possible to transform the dataset with different granuality of the time dimension in order to reveal such trends.
- The lag scales simply show a trend of decreasing price at a somewhat constant rate crossing into decreasing price at a constant rate. The crossover occurs significantly at the point of the spike in price. Most of the PACF lag is within the range for being attributed to white noi

Evaluation of *wool*

```
wool <- ts(woolyrnq, start=1965, frequency=4)
#describe(wool)
head(wool)

##      Qtr1 Qtr2 Qtr3 Qtr4
## 1965 6172 6709 6633 6660
## 1966 6786 6800

plot(wool)
```



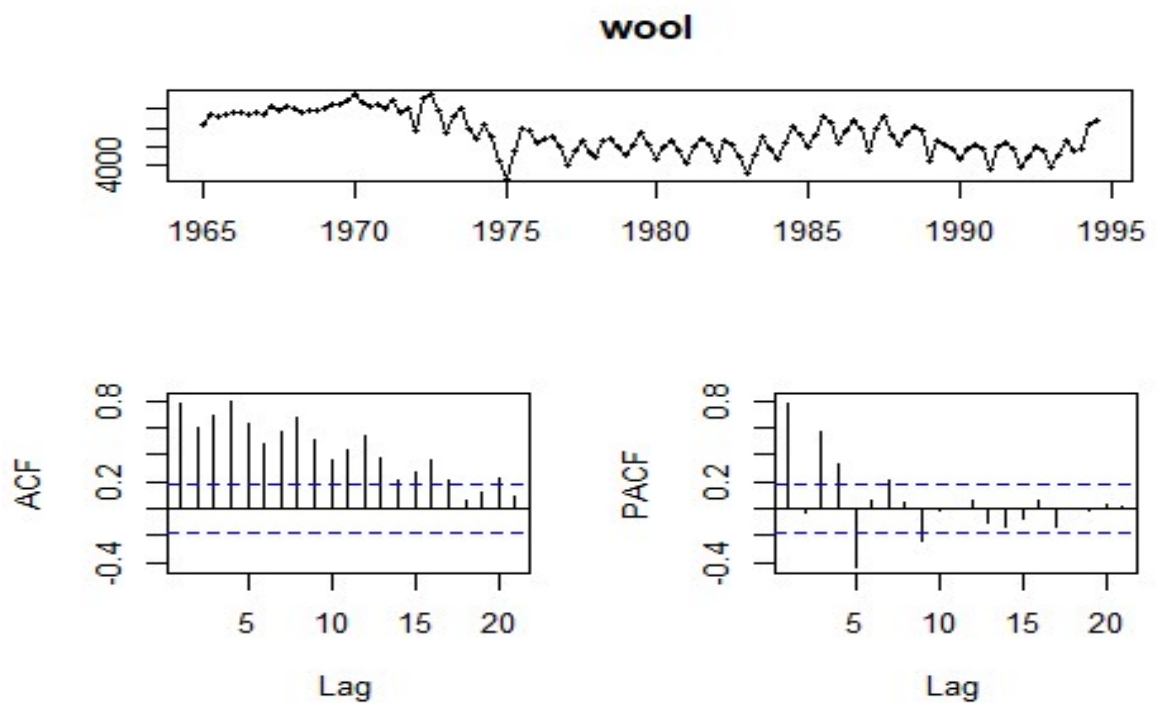

```
seasonplot(wool)
```




```
help(wool)
```

```
## No documentation for 'wool' in specified packages and libraries:
## you could try '??wool'
```

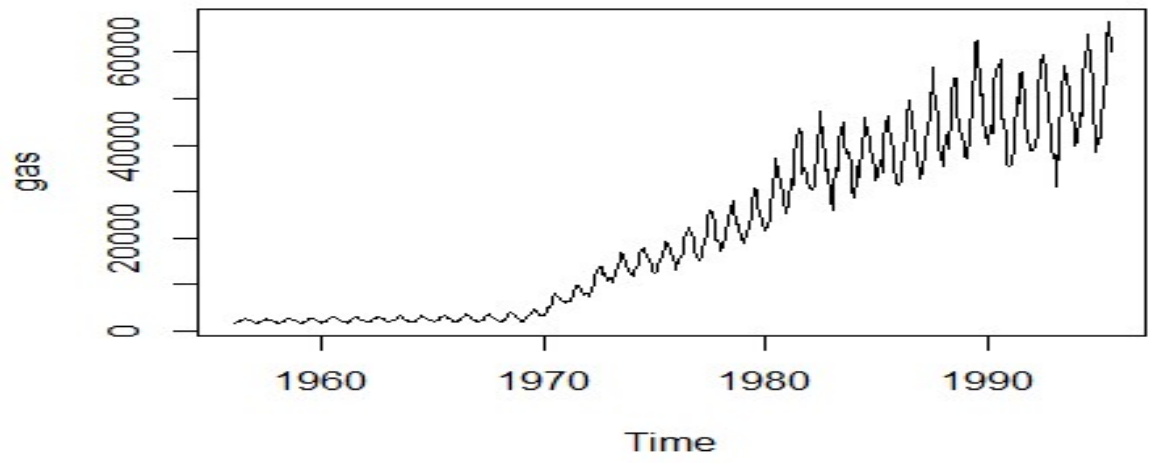
```
tsdisplay(wool)
```



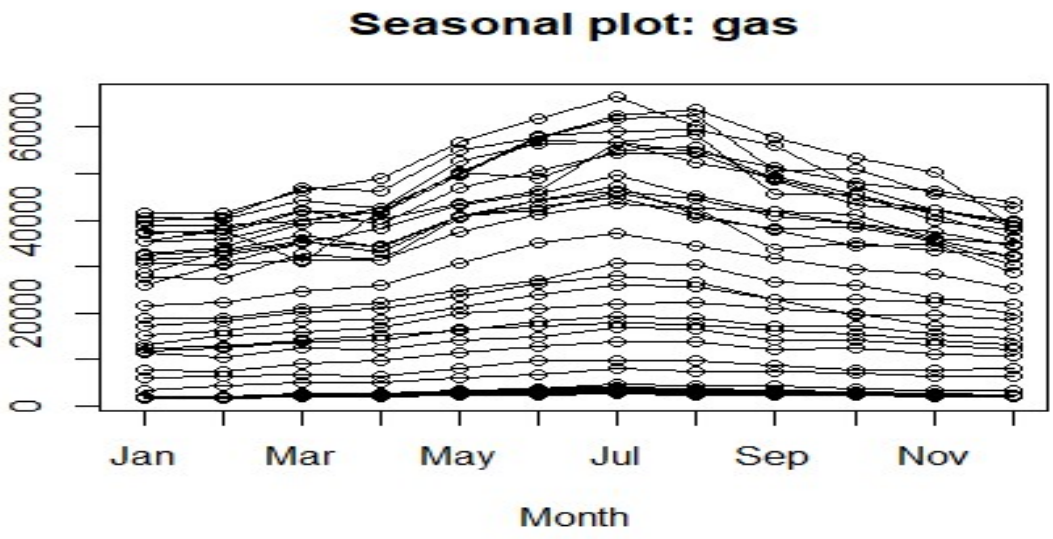
- The `wool` dataset is not structured as a timeseries object but this is easily remedied. It is clear from inspecting the data that the time dimension is quarterly beginning in 1965. The ts (timeseries) function converts the data so seasonality or year-over-year analysis is possible. It isn't clear what the `measure` dimension is capturing but perhaps it represents unit price or amount produced. In any case the trend shows an inital rise during the late 1960's followed by a greater drop during the 1970's and early 1980's, some fluctuation in the later 1980's early 1990's followed by a spike in the mid 1990's. The seasonality plots show peaks in Q3 and nadirs in Q1.
- The lag plots show slight diminishing ACF trend with cyclical seasonal fluctuation and cyclical PACF with a diminishing magnitude.

Evaluation of *gas*

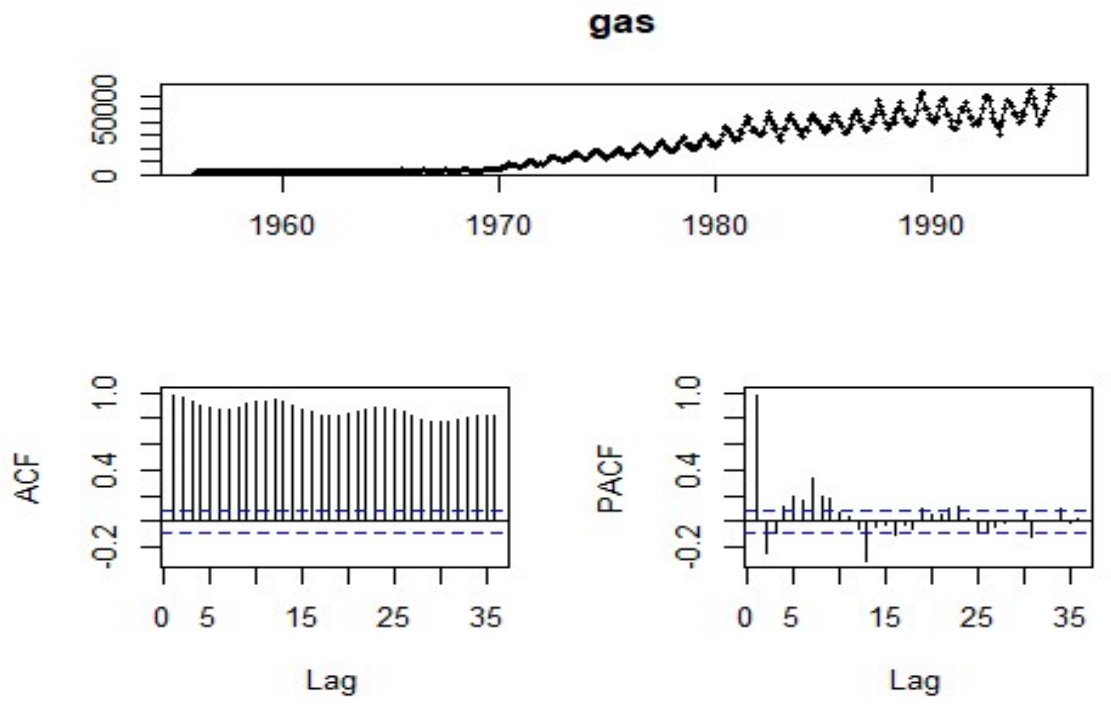
```
help(gas)
#describe(gas)
head(gas)

##       Jan  Feb  Mar  Apr  May  Jun
## 1956 1709 1646 1794 1878 2173 2321

plot(gas)
```




```
seasonplot(gas)
```



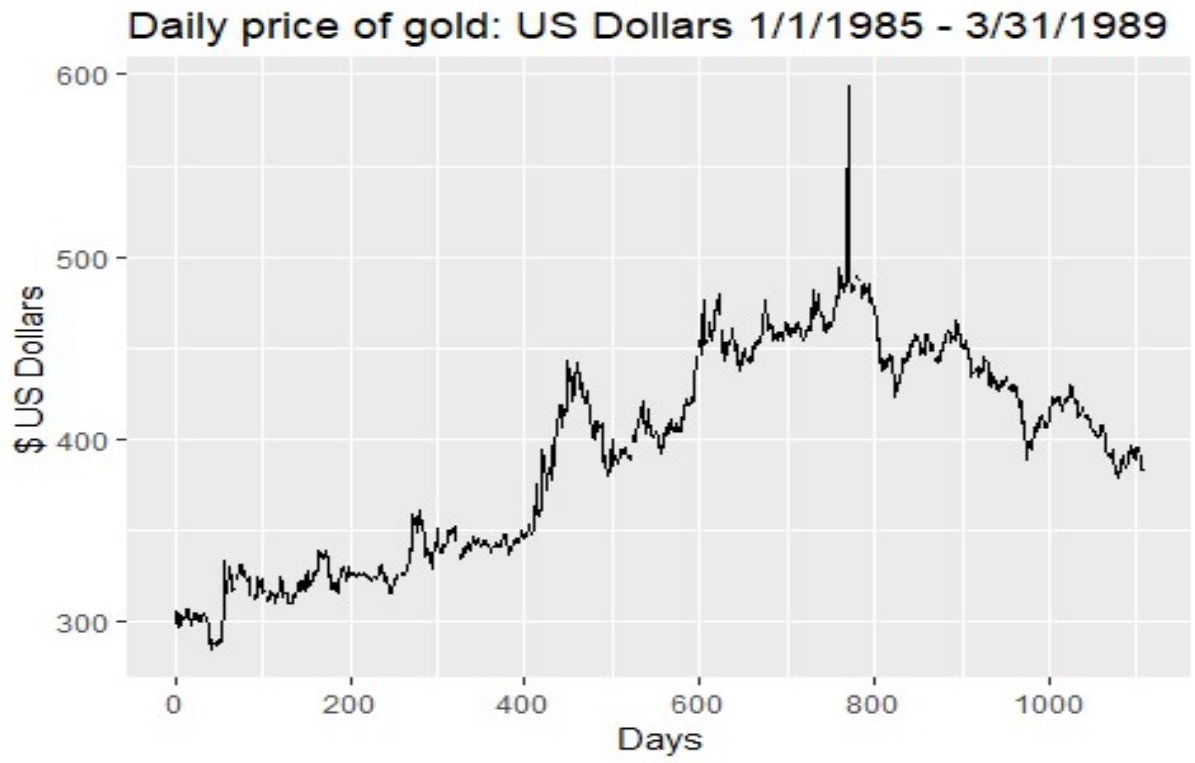

```
tsdisplay(gas)
```



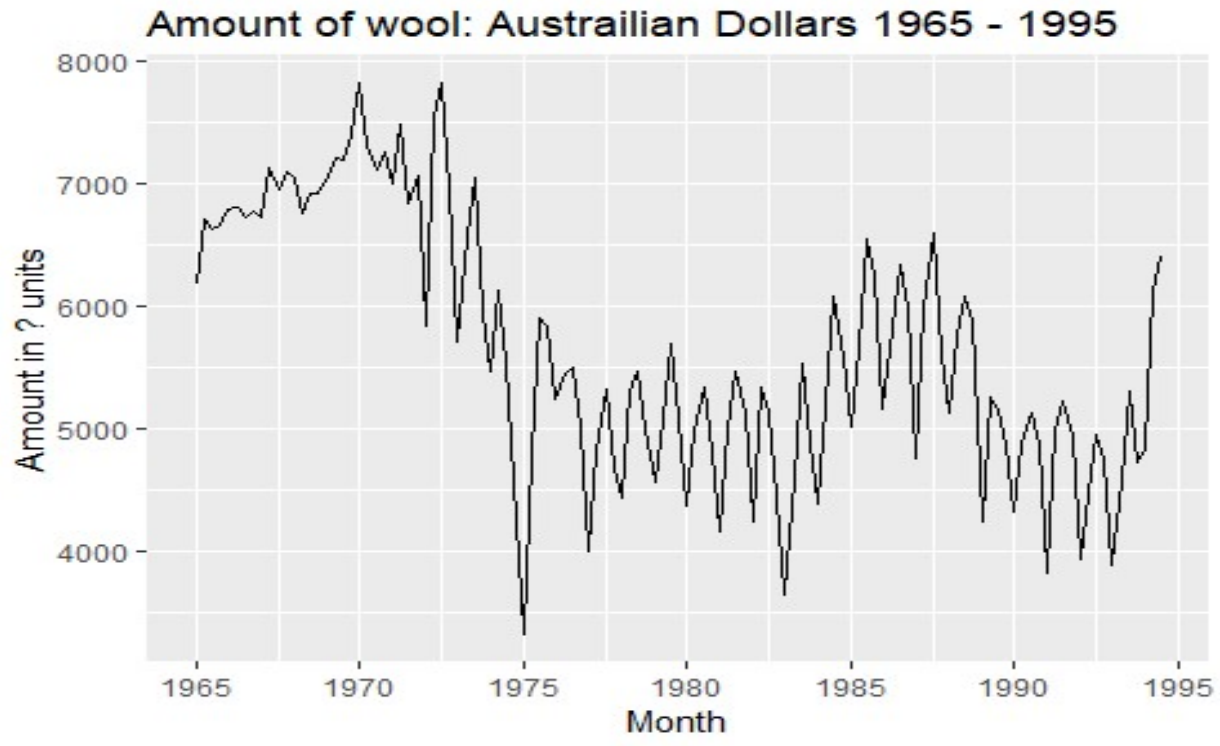
- The gas dataset is structured as a timeseries object. The description notes that the dimensions being compared are time months 1956- 1995 against production of Austrailian gas in 10K increments of unit volume.
- Based on the plot, the production of gas is flat up until the 1970s with slight seasonal fluctuation. From 1970 onward however, the trend shows increasing production at an increasing rate coupled with an increase in seasonal fluctuation.
- The seasonal plot shows peak production during July–winter in Austrailia–with a nadir during December and January–summer month.
- The lag plots show slight diminishing ACF trend with cyclical seasonal fluctuation and cyclical PACF with a diminishing magnitude. The magnitude becomes small enough toward the end that it could be attributed to white noise.

a. Use autoplot() to plot each of these in separate plots.

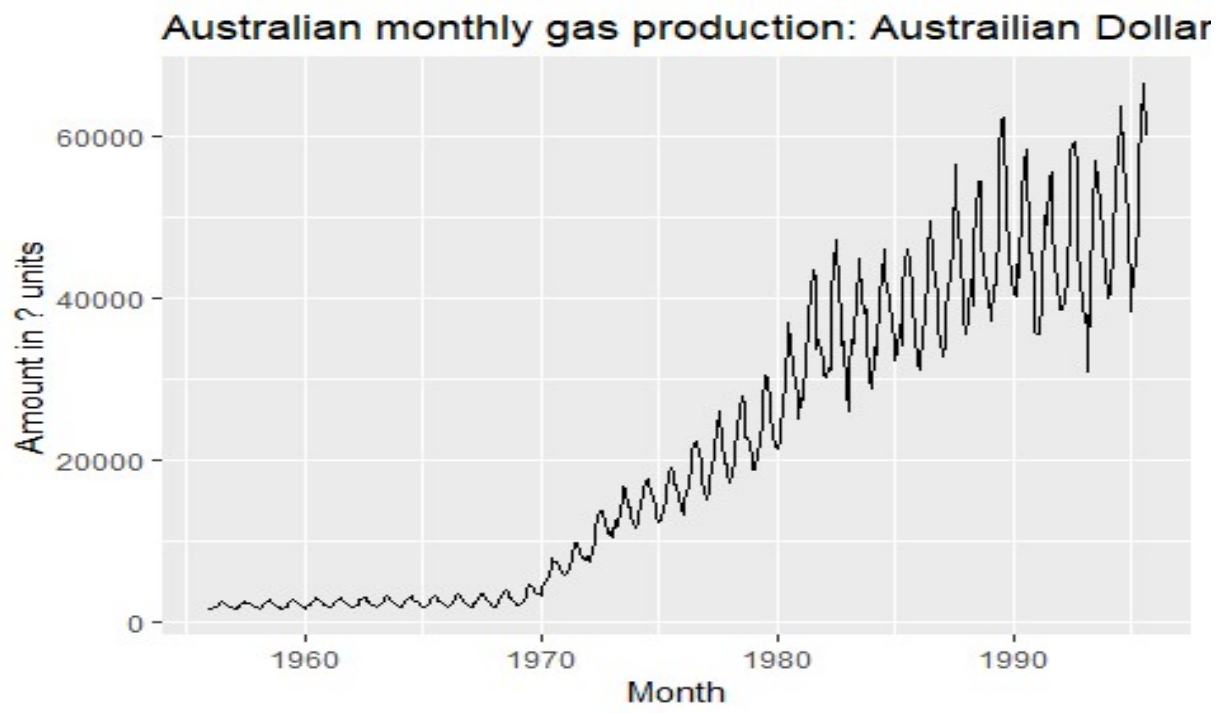
```
autoplot(gold) +
ggtitle("Daily price of gold: US Dollars 1/1/1985 - 3/31/1989") +
xlab("Days") +
ylab("$ US Dollars")
```




```
autoplot(wool) +
ggtitle("Amount of wool: Austrailian Dollars 1965 - 1995") +
xlab("Month") +
ylab("Amount in ? units")
```



```
  autoplot(gas) +
  ggtitle("Australian monthly gas production: Australian Dollar 1956- 1995.")
+
  xlab("Month") +
  ylab("Amount in ? units")
```



b. What is the frequency of each series? Hint: apply the frequency() function.

```
freq.df <- data.frame(frequency(gold),frequency(woolyrnq),frequency(gas))
names(freq.df) <- c("Frequency Gold", "Frequency Woolyrnq", "Frequency Gas")
freq.df
```

```
##   Frequency Gold Frequency Woolyrnq Frequency Gas
## 1              1                  4            12
```

- Gold:For some reasons, the function indicates that our data is annual basis, while in reality it is on daily basis. It seems to be a glitch.
- Woolyrnq: the function correctly shows that our data is on quaterly basis.
- Gas: Gas data is provided on monthly basis.

c. Use which.max() to spot the outlier in the gold series. Which observation was it?

```
paste0("Maximum Gold (Outlier Detection): ", which.max(gold))
```

```
## [1] "Maximum Gold (Outlier Detection): 770"
```

- Spike for gold prices has happened on day 770.

2.3 Download some monthly Australian retail data from the book website. These represent retail sales in various categories for different Australian states, and are stored in a MS-Excel file.

a. You can read the data into R with the following script:

```
temp = tempfile(fileext = ".xlsx")
dataURL <- "https://otexts.com/fpp2/extrafiles/retail.xlsx"
download.file(dataURL, destfile=temp, mode='wb')

retaildata <- readxl::read_excel(temp, sheet =1, skip =1)
kable(head(retaildata[1:6,1:6]))
```

Series ID	A3349335T	A3349627V	A3349338X	A3349398A	A3349468W
1982-04-01	303.1	41.7	63.9	408.7	65.8
1982-05-01	297.8	43.1	64.0	404.9	65.8
1982-06-01	298.0	40.3	62.7	401.0	62.3
1982-07-01	307.9	40.9	65.6	414.4	68.2
1982-08-01	299.2	42.1	62.6	403.8	66.0
1982-09-01	305.4	42.0	64.4	411.8	62.3

b. Select one of the time series as follows (but replace the column name with your own chosen column):

COlumn "A3349337W": "Turnover ; New South Wales ; Hardware, building and garden supplies retailing"

```
myts <- ts(retaildata[,"A3349337W"], frequency=12, start=c(1982,4))
myts
```

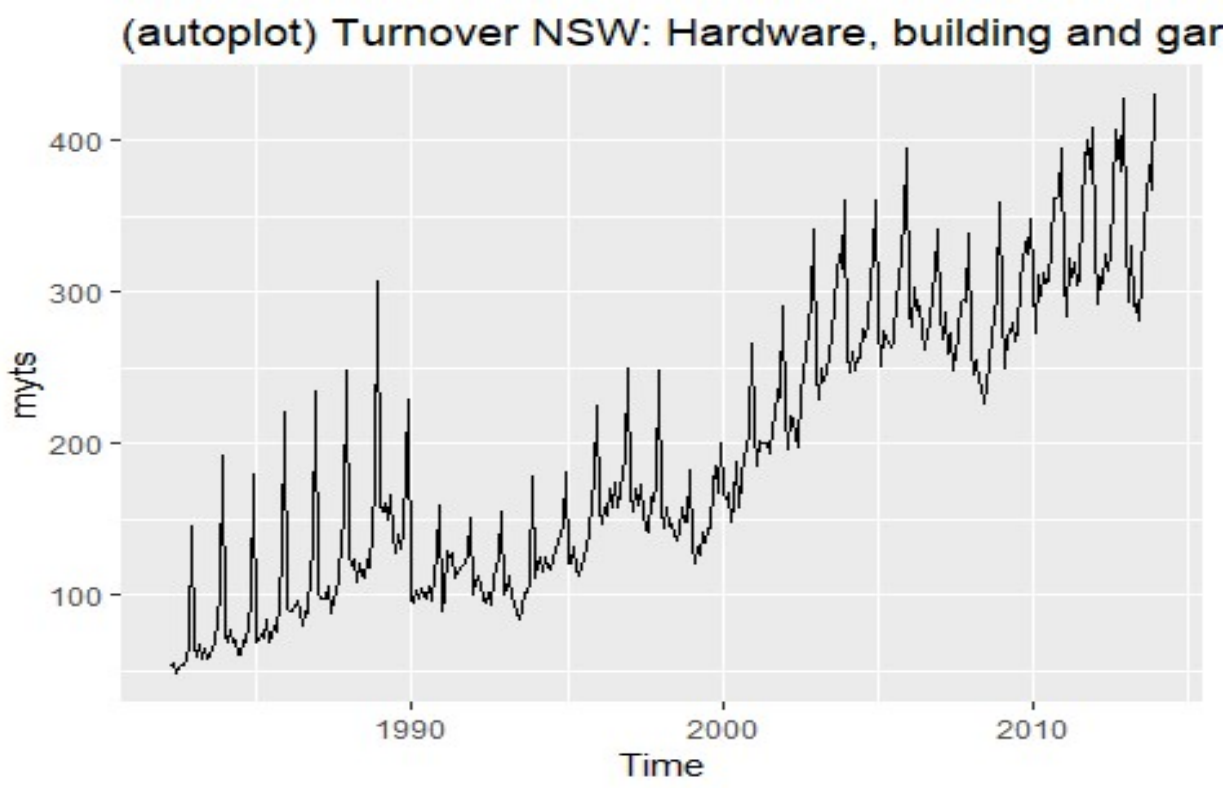
```
##        Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov
## 1982                     53.6  55.4  48.4  52.1  54.2  53.6  58.0  67.2
## 1983  66.6  59.2  67.3  57.7  64.9  58.6  58.8  64.8  68.7  84.1 101.2
## 1984  73.7  69.6  77.7  68.5  70.0  60.5  60.2  70.0  69.5  81.5  96.5
## 1985  69.4  69.8  74.1  71.9  83.6  68.8  71.8  79.4  76.0  97.0 126.8
## 1986  90.3  89.8  89.6  91.9  96.0  89.3  79.4  89.1  88.1 116.8 128.6
## 1987 103.9  97.3  97.9  97.2 106.5  88.2  97.7 100.2 110.8 137.3 150.5
## 1988 126.6 119.4 123.6 108.8 121.0 113.9 110.9 124.3 118.5 143.9 172.1
## 1989 160.7 155.2 161.0 149.3 165.6 140.1 128.2 140.4 130.2 143.3 185.3
## 1990  96.4  95.0 103.8  97.1 104.6 100.7  98.2 106.6  96.7 113.3 126.2
## 1991  89.1  99.6 129.0 125.6 127.3 111.7 114.1 118.0 119.6 121.5 128.5
## 1992 100.1 108.2 113.2 108.0  98.2  95.2 101.4  93.5 112.0 118.9 125.7
## 1993 100.7 102.8 113.5  99.2  95.4  89.3  84.4  91.1 102.2 101.4 108.5
## 1994 111.0 121.4 125.6 116.2 125.1 119.1 117.5 123.8 134.5 141.0 145.2
## 1995 120.8 121.0 132.6 116.3 113.2 120.2 124.3 134.0 140.6 163.7 176.2
## 1996 157.5 147.7 158.1 152.4 171.0 158.0 174.0 157.5 167.0 181.0 189.6
## 1997 168.0 154.9 169.9 159.8 172.7 154.1 144.9 141.3 164.3 162.7 172.8
## 1998 157.0 145.0 158.6 145.9 146.8 140.2 135.8 141.7 158.7 148.4 148.0
## 1999 133.1 120.5 132.2 126.0 141.0 135.0 143.7 144.4 171.7 185.5 167.9
```

```
## 2000 169.7 163.2 167.6 148.7 161.4 188.5 158.3 174.5 193.2 194.5 209.7
## 2001 209.6 185.2 202.2 200.0 200.3 200.3 193.6 211.4 218.2 236.3 230.6
## 2002 219.9 196.6 218.7 216.8 205.5 198.2 233.9 246.2 259.8 277.3 294.3
## 2003 247.0 229.3 250.3 241.6 247.0 258.7 271.3 291.1 312.7 324.6 315.2
## 2004 258.9 246.5 260.9 249.0 256.5 257.4 275.4 269.8 279.8 307.3 323.9
## 2005 281.8 250.6 274.1 270.3 268.2 264.0 266.9 298.6 303.1 329.4 345.6
## 2006 288.0 277.3 302.8 288.5 290.4 275.4 262.4 272.9 279.7 299.3 313.3
## 2007 286.4 268.4 286.6 260.0 273.0 248.5 259.7 272.2 293.6 294.9 294.3
## 2008 263.0 246.2 255.2 240.2 239.6 226.9 238.7 253.1 271.3 283.1 299.0
## 2009 289.3 249.6 272.1 272.9 279.4 267.8 273.1 307.7 318.2 334.0 325.0
## 2010 309.2 272.6 311.1 298.2 313.1 305.8 307.3 330.9 362.8 361.7 364.2
## 2011 311.6 283.7 322.2 310.8 319.5 305.1 308.9 355.6 384.9 401.1 382.1
## 2012 334.0 292.1 309.6 305.8 325.0 314.2 327.2 363.7 406.9 397.1 379.6
## 2013 340.0 293.9 330.7 290.7 291.8 281.1 309.8 344.6 360.7 384.7 367.9
##        Dec
## 1982 146.3
## 1983 192.3
## 1984 179.4
## 1985 221.2
## 1986 235.4
## 1987 248.8
## 1988 307.4
## 1989 228.9
## 1990 159.5
## 1991 151.4
## 1992 154.7
## 1993 179.0
## 1994 180.7
## 1995 225.4
## 1996 249.8
## 1997 248.7
## 1998 183.0
## 1999 200.7
## 2000 266.3
## 2001 291.0
## 2002 341.9
## 2003 360.8
## 2004 361.1
## 2005 395.2
## 2006 341.6
## 2007 339.3
## 2008 360.2
## 2009 348.9
## 2010 395.4
## 2011 409.0
## 2012 428.0
## 2013 430.7
```

c. Explore your chosen retail time series using the following functions

-autoplot() A trend exists when there is a long-term increase or decrease in the data. As per the below (autoplot) there seems to be a general upward trend in the data

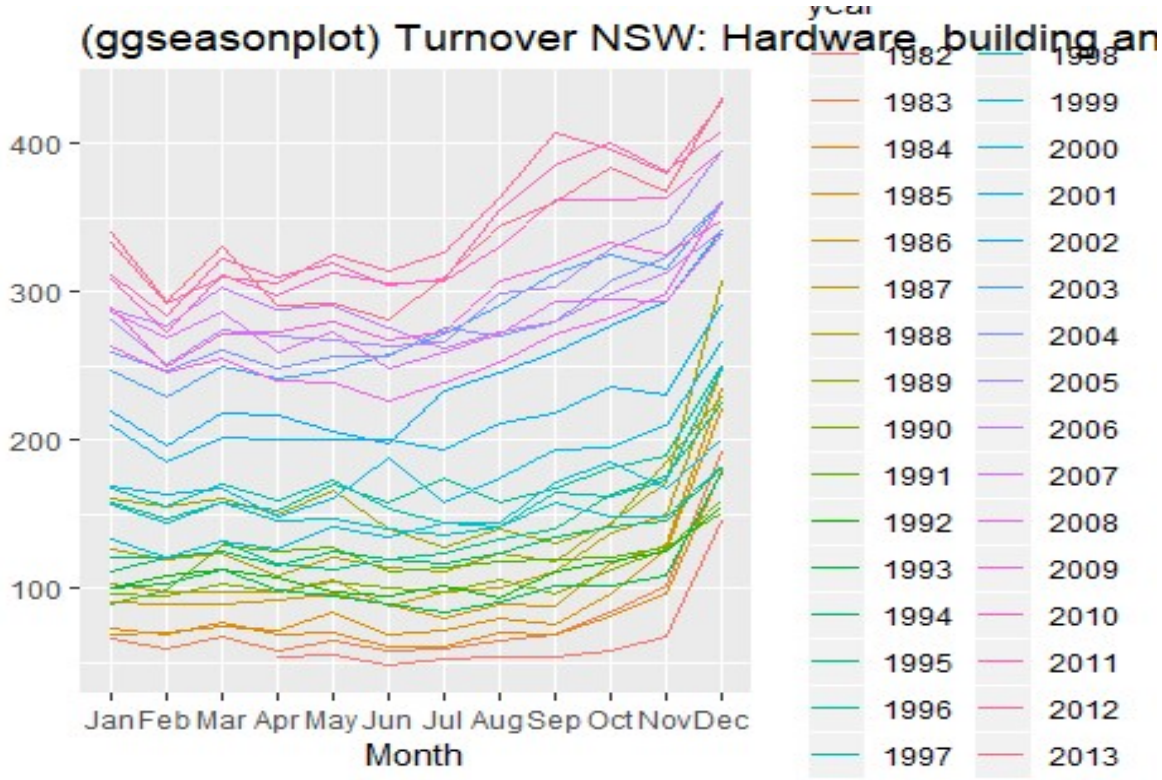
```
autoplot<- autoplot(myts) + ggtitle("(autoplot) Turnover NSW: Hardware, build
ing and garden supplies retailing")
autoplot
```



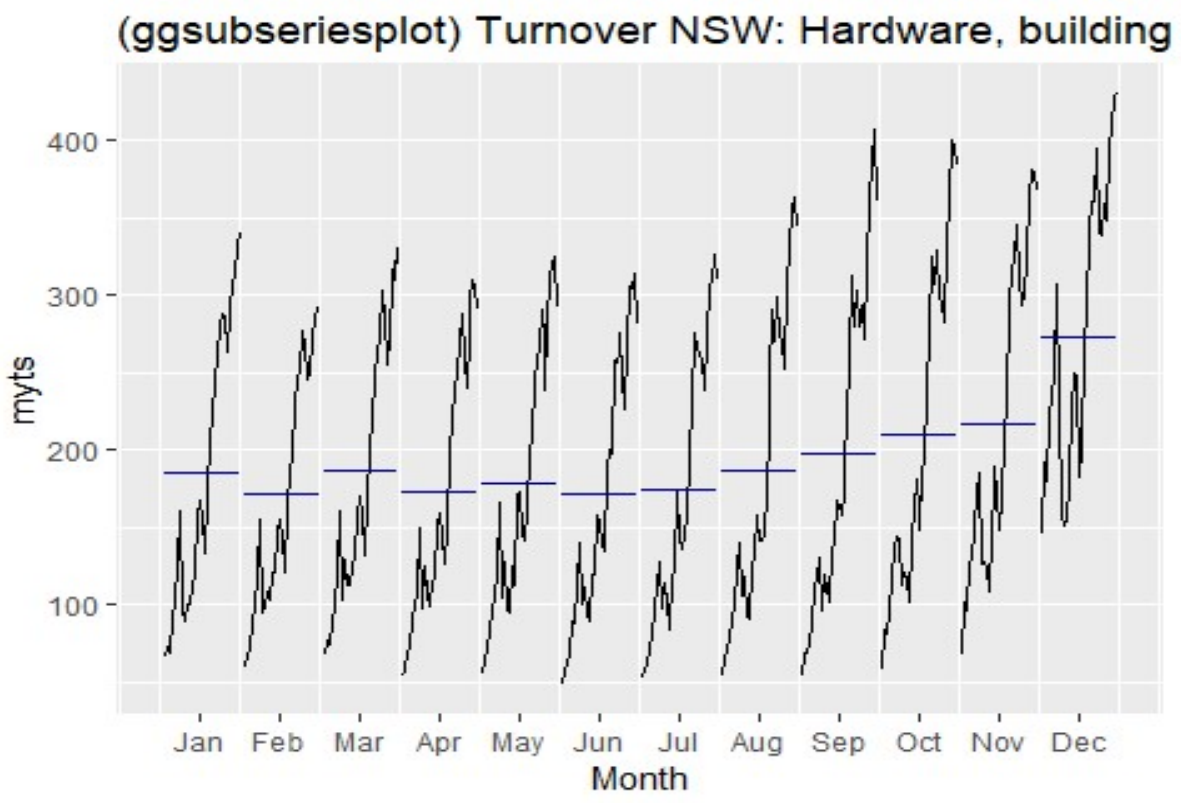
- Seasonality: A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. The seasonal plots are similar to a time plot except that the data are plotted against the individual "seasons" in which the data were observed. Below, the turnover in NSW for Hardware, building and garden supplies retailing is plotted using the seasonal plots to test seasonality. The data shows that the turnover is higher in the September, October, November and December months. This makes sense as Australia spring season starts in September and ends in late December.This is especially exemplified in the lag plots with the largest lags 12 months apart and the polar seasonplot
    - ggseasonplot(): It seems our data tend to increase slightly in December. February, the shortest month, seems to dip a little bit
    - ggsubseriesplot(): Again, we see that December is the highest month and February is the lowest
    - gglagplot(): If we look at Lag 12 we can see very strong indication of autocorrelation.

- ggAcf(): Consistent decrease due to trend and very slight "scalloped" shape due to slight seasonality.

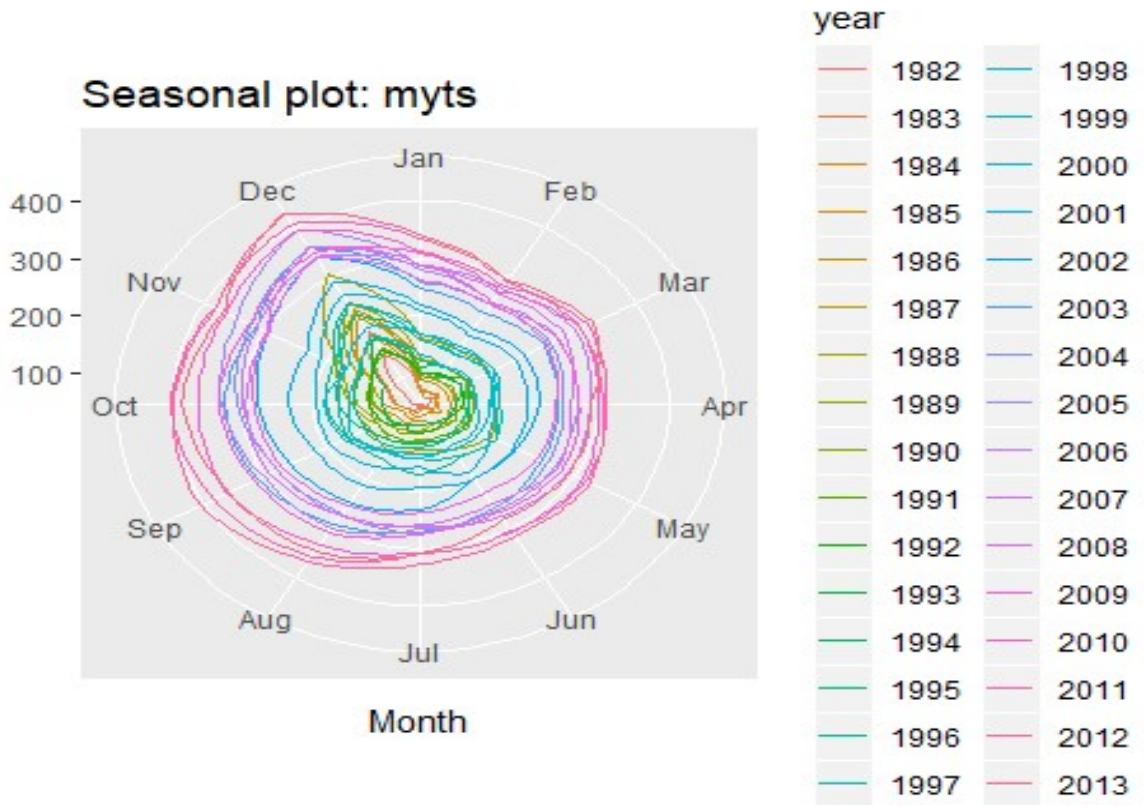
```
#seasonality

ggseasonplot<- ggseasonplot(myts)+ ggtitle("(ggseasonplot) Turnover NSW: Hard
ware, building and garden supplies retailing")
ggsubseriesplot<- ggsubseriesplot(myts)+ ggtitle("(ggsubseriesplot) Turnover
NSW: Hardware, building and garden supplies retailing")

ggseasonplotpolar<-ggseasonplot(myts, polar=TRUE)
gglagplot<- gglagplot(myts)+ ggtitle("(gglagplot) Turnover NSW: Hardware, bui
lding and garden supplies retailing")
ggAcf<- ggAcf(myts)+ ggtitle("(ggAcf) Turnover NSW: Hardware, building and ga
rden supplies retailing")

ggseasonplot
```



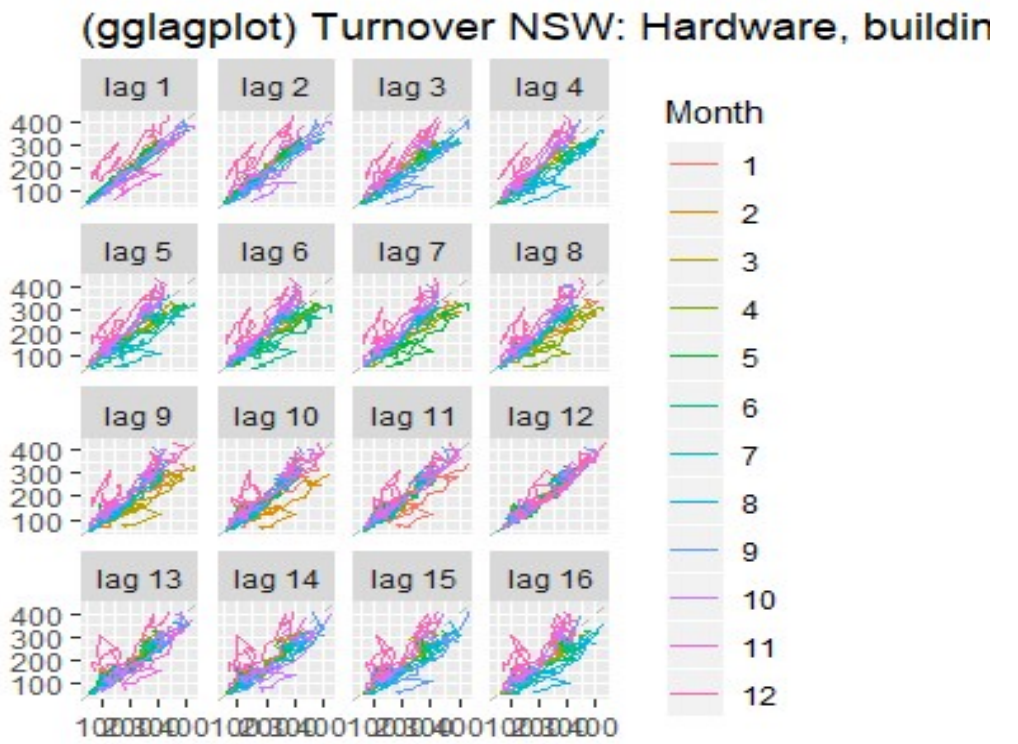

```
ggsubseriesplot
```



ggseasonplotpolar



gglagplot



ggAcf