data_624_homeworkone

Kevin Kirby

2024-09-04

Homework One for Fall 2024 DATA 624 at CUNY School of Professional Studies

The below are answers to exercises 2.1, 2.2, 2.3, 2.4, 2.5 and 2.8 from section 2.10 of the [Forecasting: Principles and Practice (3rd ed)] (https://otexts.com/fpp3/graphics-exercises.html). This is being completed for homework one of the DATA 624 class "Predictive Analytics."

Exercises

2.1 Exploring Time Series

The exercise states: "Explore the following four time series: Bricks from aus_production, Lynx from pelt, Close from gafa_stock, and Demand from vic_elec.

I have installed a few libraries needed for these exercises, which I'm going to load at the top for efficiency:

tsible and tsibbledata: the packages where the data lives ggplot2: common tidyverse plotting library dplyr: common data manipulation package fabletools: plotting for tsibble objects

```
library(fpp3)
```

```
## Registered S3 method overwritten by 'tsibble':
##   method               from
##   as_tibble.grouped_df dplyr
```

```
## -- Attaching packages --------------------------------------- fpp3 1.0.0 --
```

```
## v tibble      3.2.1     v tsibble     1.1.5
## v dplyr       1.1.4     v tsibbledata 0.4.1
## v tidyr       1.3.1     v feasts      0.3.2
## v lubridate   1.9.3     v fable       0.3.4
## v ggplot2     3.5.1     v fabletools  0.4.2
```

```
## -- Conflicts ------------------------------------------ fpp3_conflicts --
## x lubridate::date()    masks base::date()
## x dplyr::filter()      masks stats::filter()
## x tsibble::intersect() masks base::intersect()
## x tsibble::interval()  masks lubridate::interval()
## x dplyr::lag()         masks stats::lag()
## x tsibble::setdiff()   masks base::setdiff()
## x tsibble::union()     masks base::union()
```

```
library(tsibble)
library(tsibbledata)
library(ggplot2)
library(dplyr)
```

```
library(fabletools)
library(readr)
data("aus_production")
data("pelt")
data("gafa_stock")
data("vic_elec")
data("us_employment")
data("PBS")
data("us_gasoline")
```

For each time series, the exercise asks me to answer the following questions:

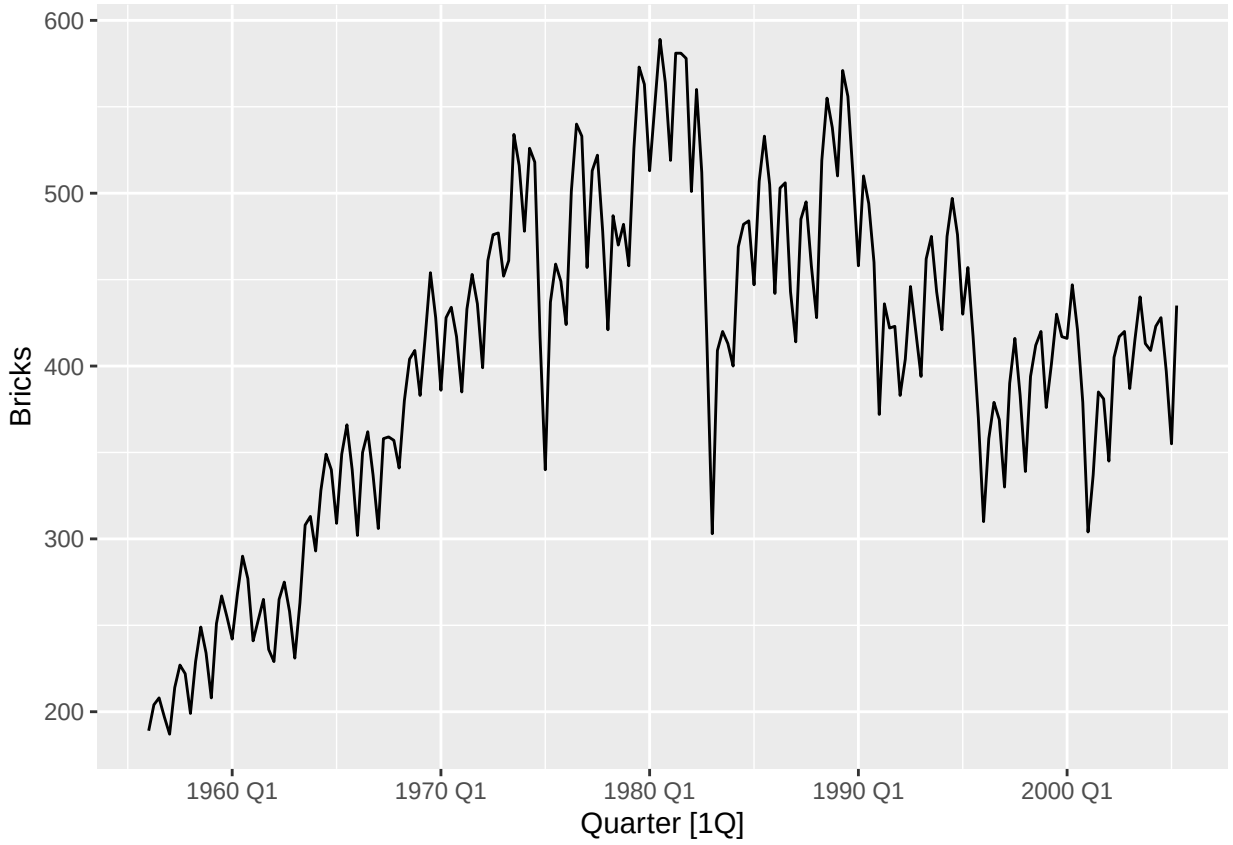
What is the time interval of each series? Use autoplot() to produce a time plot of each series.

For the last time series, it asks me to modify the axis labels and title.

Bricks from aus_production: Time interval: every three months (a calendar year quarter) covering 1956 Q1 to 2010 Q2

Using autoplot() to produce a timeplot:

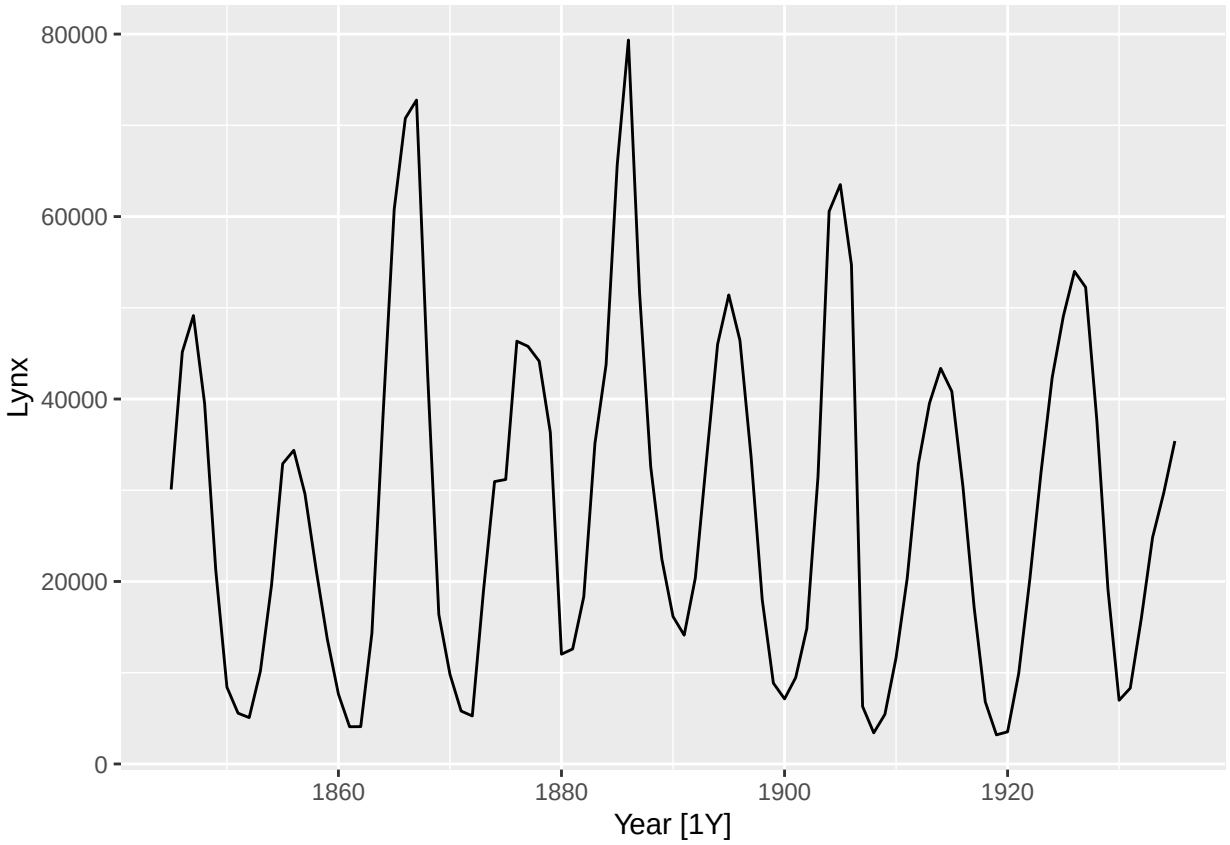
```
aus_production |>
  filter(!is.na(Bricks)) |>
  autoplot(Bricks)
```



Lynx from pelt: Time interval: 12 months (one calendar year) from 1845 to 1935

Using autoplot() to produce a timeplot:

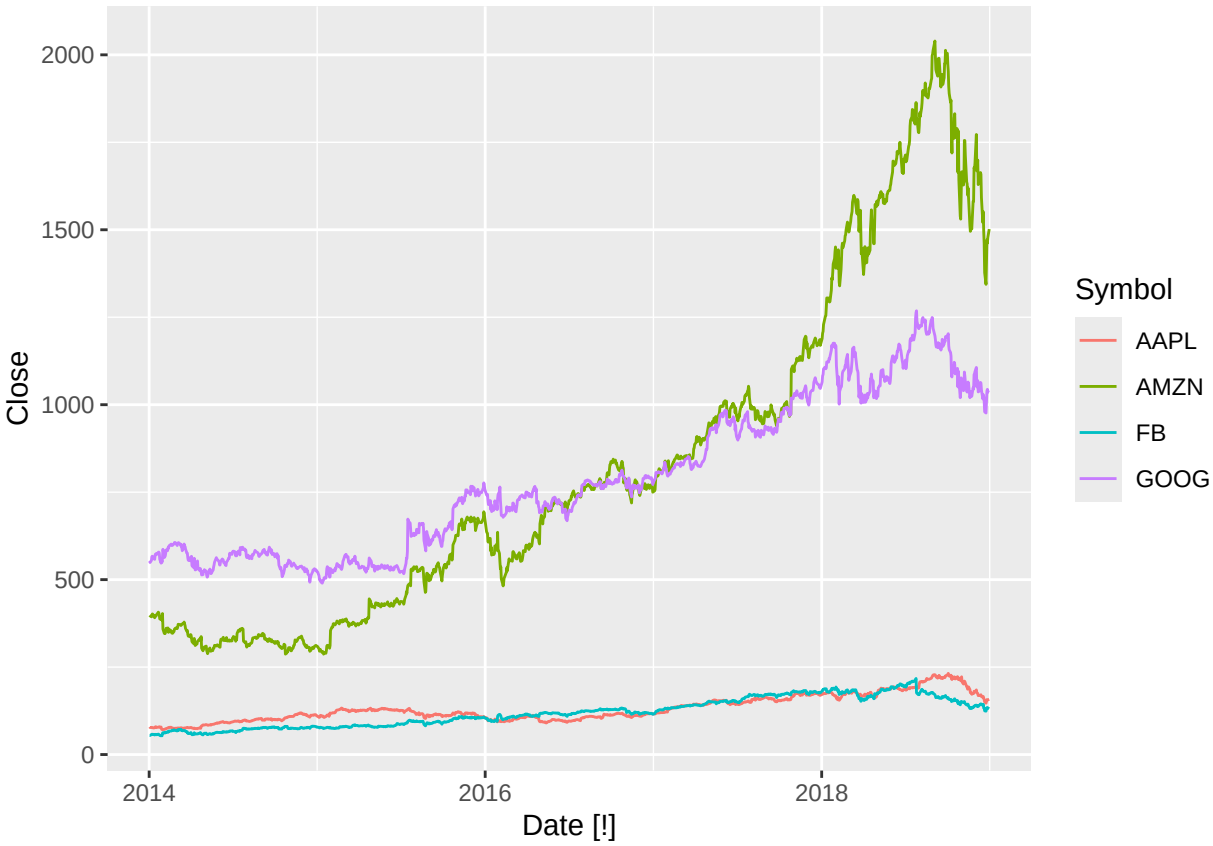
```
pelt |>
  filter(!is.na(Lynx)) |>
  autoplot(Lynx)
```



Close from gafa_stock Time interval: one day interval from 01/02/2014 to 12/31/2018

Using autoplot() to produce a timeplot:

```
gafa_stock |>
  filter(!is.na(Close)) |>
  autoplot(Close)
```



Demand from vic_elec Time interval: every 30 minutes from 1/1/2012 at 12:00am AEDT to 23:30 AEDT on 12/31/2014:

For this autoplot, I will modify the title and axis labels as part of the autoplot function:

"'{r-electricity-demand} vic_elec |> filter(!is.na(Demand)) |> autoplot(Demand) + labs(title = "Electricity Demand for Victoria, Australia - 30 Minute Intervals", x = "Date & Time", y = "Operational Demand")

```
### 2.2 - Using Filter on gafa_stock

Exercise 2.2 asks me to "use filter() to find what days corresponded to the peak closing price for each


```{r-gafa-filter}

gafa_stock %>%
  group_by(Symbol) %>%
  filter(Close == max(Close, na.rm = TRUE))
```

This code will tell you the following dates have the highest close prices for each stock:

AAPL 2018-10-03 AMZN 2018-09-04 FB 2018-07-25 GOOG 2018-07-26

2.3 Working with Tute1 dataset:

Exercise 2.3 has three components, each of which I'll discuss as I do them. All of them require working with the tute1 dataset that's provided on the bookset.
```

2.3.a - For A, I have to use the read_csv function from the tidyverse package readr to read and view the tute1.csv file that's provided by the book website.

I loaded the file into a Google Cloud Platform bucket that's designed for my CUNY work. I set it up so I can permission files at the file level, allowing for public access as needed while protecting the overall security of the bucket. When working programmatically in python, I use authewnticated user access to bring in the file.

```
gcp_tute_url <- "https://storage.googleapis.com/cuny_files/data_624_forecasting/weektwo_homework_one/tu
gcp_tute1 <- read_csv(url(gcp_tute_url))
```

```
## Rows: 100 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (3): Sales, AdBudget, GDP
## date (1): Quarter
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
gcp_tute1
```

```
## # A tibble: 100 x 4
##    Quarter    Sales AdBudget   GDP
##    <date>     <dbl>    <dbl> <dbl>
##  1 1981-03-01 1020.     659.  252.
##  2 1981-06-01  889.     589   291.
##  3 1981-09-01  795      512.  291.
##  4 1981-12-01 1004.     614.  292.
##  5 1982-03-01 1058.     647.  279.
##  6 1982-06-01  944.     602   254
##  7 1982-09-01  778.     531.  296.
##  8 1982-12-01  932.     608.  272.
##  9 1983-03-01  996.     638.  260.
## 10 1983-06-01  908.     582.  280.
## # i 90 more rows
```

2.3.b - Convert to Time Series: This question just asks me to convert to a time series with provided code:

"'{r-time-series-conversion} tute_timeseries <- gcp_tute1 |> mutate(Quarter = yearquarter(Quarter)) |> as_tsibble(index = Quarter)

```
#### 2.3.c - Plot Time Series
The final question asks me to plot of each of the three series in the time-converted series I made. Addi

```{r-tute_timeseries-plot}
tute_timeseries |>
  pivot_longer(-Quarter) |>
  ggplot(aes(x = Quarter, y = value, colour = name)) +
  geom_line() +
  facet_grid(name ~ ., scales = "free_y")
```

Removing facet_grid combines the three lines into one chart where the lines are stacked on top of each other. I prefer it for these three since the y-axis scales for each are similar enough and it removes the strange right hand side y-axis label for the metric, which is redundant with the legend right there.

If the y-axis scales were a lot different then you would end up with a scale warped to the largest values.
```

“‘{r-tute_timeseries-plot} tute_timeseries |> pivot_longer(-Quarter) |> ggplot(aes(x = Quarter, y = value, colour = name)) + geom_line()

2.4 - Working With USgas Package

Exercise 2.4 also three components that involves installing and working with the USgas Package:

2.4.a - Install and Load

The first component only asks me to install the package, which will naturally involve loading it as wel

```{r-install-us-gas}
install.packages("USgas")
library(USgas)
```

2.4.b - Create Tsibble The second component is to create a tsibble fromn the us_total dataset within USGas

“‘{r-create-tsibble} us_total_tsibble <- us_total %>% as_tsibble(index = year, key = state)

2.4.c - Plot New England Gas State Consumption

The final component is to plot the annual natural gas consumption for the New England region: Maine, Ve

As a proud boy from Portland, Maine, I can say we have the best maple syrup in the country. This is not

To create the plot, I will park the New England state names in a variable, filter us_total_tsibble for

```{r-new-england-glas-plot}

install.packages("scales")
library(scales)

ne_states<- c("Maine", "Vermont", "New Hampshire", "Massachusetts", "Connecticut", "Rhode Island")

ne_filter <- us_total_tsibble %>%
  filter(state %in% ne_states)

ggplot(ne_filter, aes(x = year, y = y, color = state)) +
  geom_line() +
  labs(title = "Annual Natural Gas Consumption in New England States",
       x = "Year",
       y = "Natural Gas Consumption",
       color = "State") +
      scale_y_continuous(labels = label_number(big.mark = ",")) +
  theme_minimal()
```

2.5 - Working with Tourism Data:

Exercise 2.5 has four components, which involves loading and using the tourism.xlsx file provided by the book website.

2.5.a - Download and Load Data For A, I have to use the read_excel() function from the readxl package to read and view the tourism-xlsx file that’s provided by the book website.

Readxl does not support reading from a url so I parked this one on my local computer and will read it from there

“‘{r-tourism-load} install.packages(“readxl”) library(readxl) tourism_excel <- read_excel(“/Users/kevinkirby/Desktop/github_
tourism_excel

```
#### 2.5.b - Create Tsibble

For B, I need to make a tsibble that's exactly the same as the tourism tsibble from the tsibble package

Upon inspect that package, I see the only difference is that tourism_excel has Quarter formatted as "19
```

```{r-tourism-tissble}
tourism_excel_tsibble <- tourism_excel |>
    mutate(Quarter = yearquarter(Quarter)) |>
    as_tsibble(index = Quarter, key = c(Region, Purpose))
tourism_excel_tsibble
```

To check for differences, I can use semi_join() to perform a filtered join where there are only matches between the two, removing anything that doesn’t have a match.

“‘{r-excel-tsibble-match} excel_tsibble_match <- semi_join(tourism_excel_tsibble, tourism, by = c(“Region”, “Purpose”, “Quarter”, “Trips”))

```
#### 2.5.c - Average Maximum Overnight Trips

The third component is to "find what combination of Region and Purpose had the maximum number of overni

The answer will be:
 Region    Purpose  Quarter avg_trips
Melbourne Visiting 2017 Q4      985.
```

```{r-trip-averages}

rp_trips <- excel_tsibble_match %>%
  group_by(Region, Purpose) %>%
  summarise(avg_trips = mean(Trips, na.rm = TRUE)) %>%
  ungroup()
trip_avgs

rp_trips %>%
  filter(avg_trips == max(avg_trips))
```

2.8 - Using Graph Functions for EDA

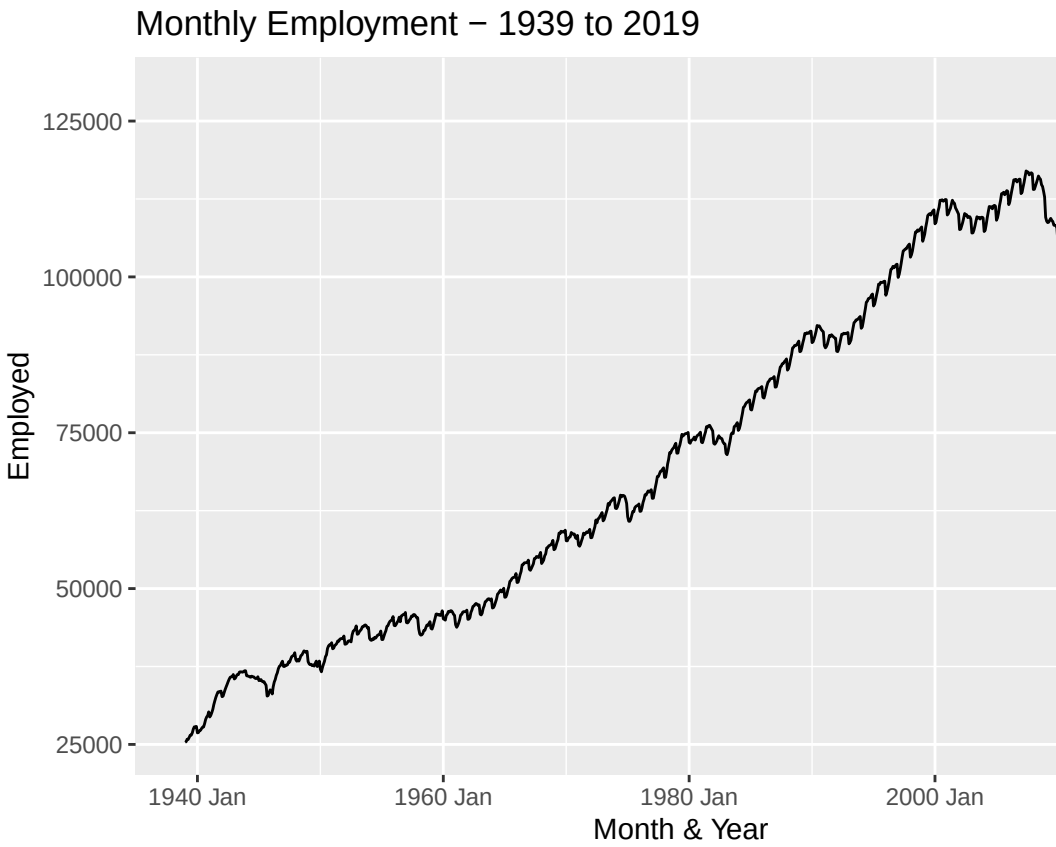
The final homework exercise is to use the following graphics functions: * autoplot() * gg_season() * gg_subseries() * gg_lag() * ACF()

With these functions, I will explore the following time series: * “Total Private” Employed from us_employment *This requires installing and loading the “fpp3” package, which is the package of the textbook authors.* Bricks from aus_production * Loaded at very start * Hare from pelt * Loaded at very start * “H02” Cost from PBS *PBS is also in the fpp3 package* Barrels from us_gasoline

I will answer the following questions from the exercise under each series of graphs by dataset:

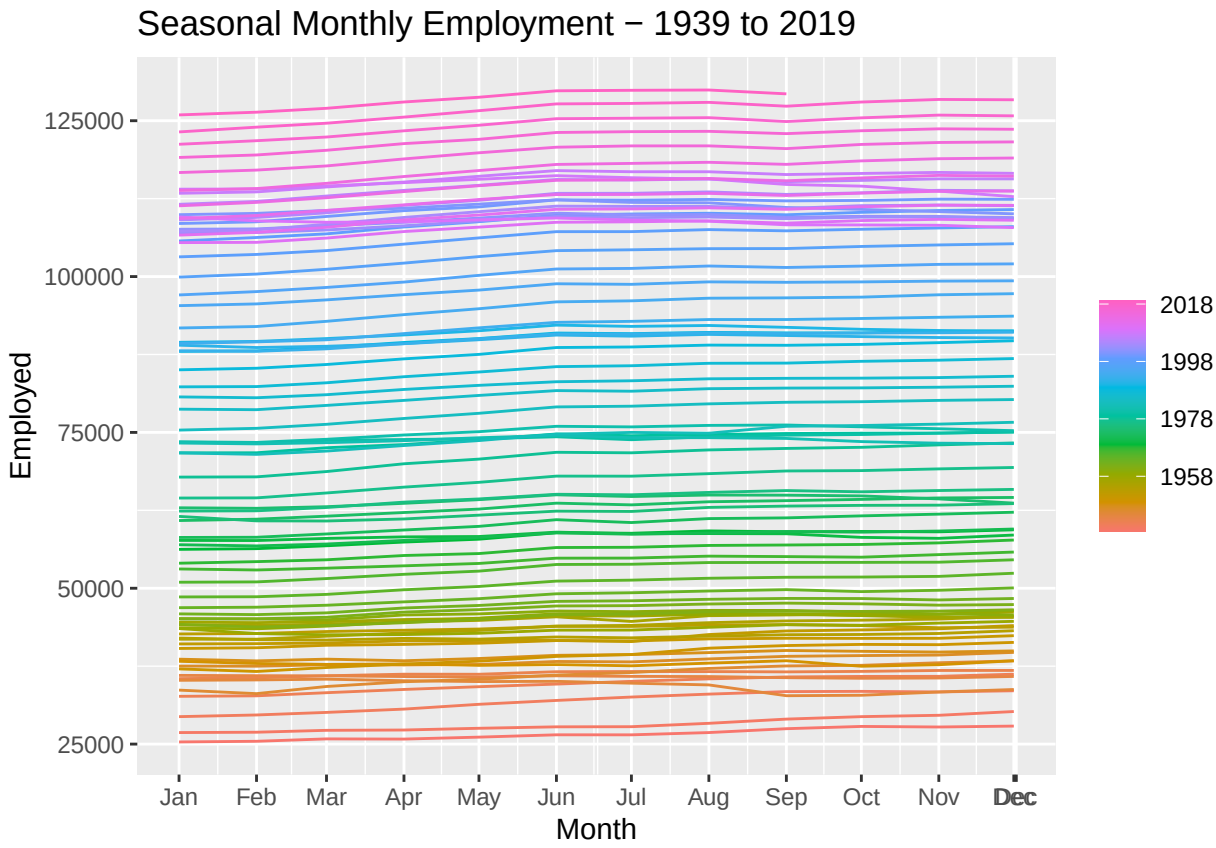
- Can you spot any seasonality, cyclicity and trend?
- What do you learn about the series?
- What can you say about the seasonal patterns?
- Can you identify any unusual years?

```
library(tidyr)

data("us_employment")
total_private <- us_employment %>%
  filter(Title == "Total Private")

total_private %>%
  drop_na(Employed) %>%
  autoplot(Employed) +
  labs(title = "Monthly Employment - 1939 to 2019",
  x = "Month & Year",
  y = "Employed")
```

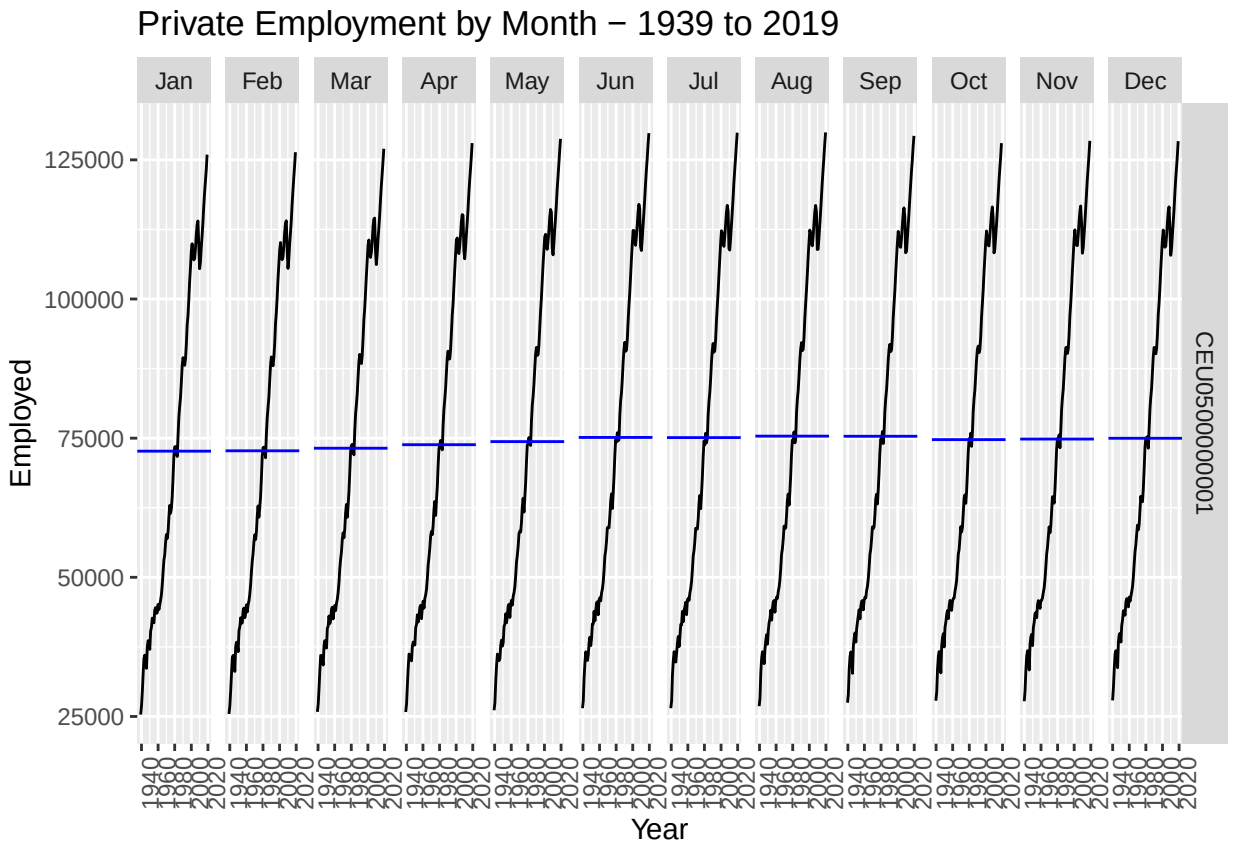


Charting Total Private Data

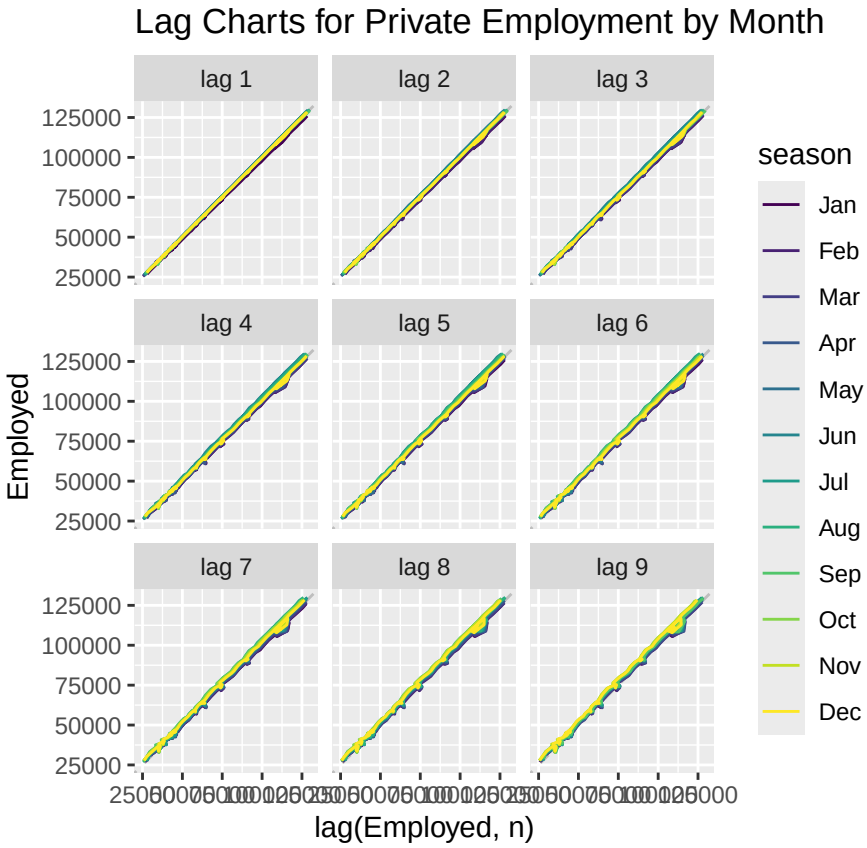
```
total_private %>%
  drop_na(Employed) %>%
  gg_season(Employed) +
  labs(title = "Seasonal Monthly Employment - 1939 to 2019",
  x = "Month",
  y = "Employed")
```



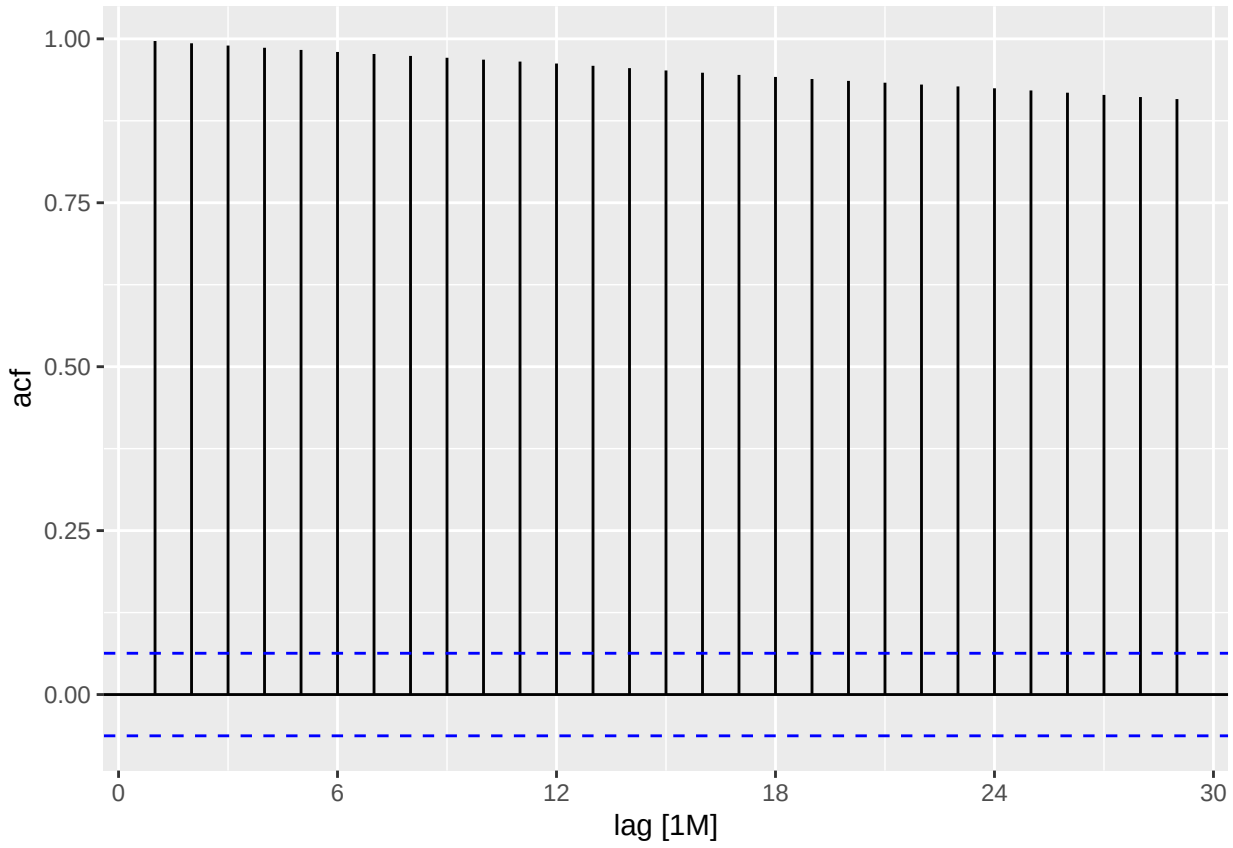

```
total_private %>%
  drop_na(Employed) %>%
  gg_subseries(Employed) +
  labs(title = "Private Employment by Month - 1939 to 2019",
  x = "Year",
  y = "Employed")
```




```
total_private %>%
  drop_na(Employed) %>%
  gg_lag(Employed) +
  labs(title = "Lag Charts for Private Employment by Month")
```




```
total_private %>%
  drop_na(Employed) %>%
  ACF(Employed) %>% autoplot()
```

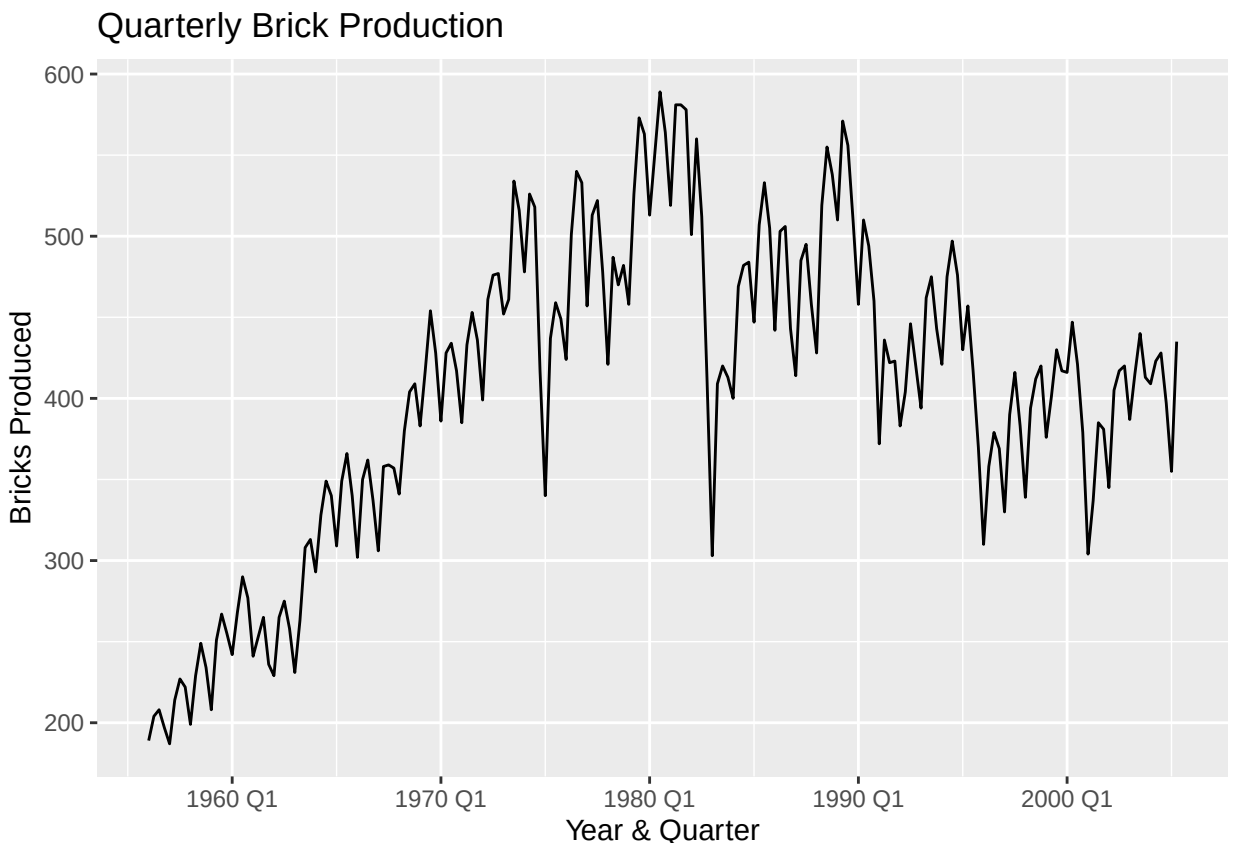


- Can you spot any seasonality, cyclicity and trend?
  - There is a strong upward trends overtime in overall private employment. However, without normalizing to account for overall population growth, this trend should be taken with a grain of salt. The strong linear relationship demonstrated in the lag charts shows strong seasonality, which the lag not impacting the relationship.
  - There are strong positive autocorrelation values from the ACF chart, showing previous employment levels will strong indicate future employment levels. This could become the basis for a forecast
- What do you learn about the series?
  - Raw private employment has continued to steadily increase over time, with only major drops around the recession periods
  - Past employment levels are a strong basis for future employment level predictions, even within a month
- What can you say about the seasonal patterns?
  - The employment for the previous season (in this case, a month in a previous year) will have a strong impact on the future employment levels
- Can you identify any unusual years?
  - Not sure unusual is the word I would use but the financial crisis of 2008-2010 really stands out in these charts and, when put in the context of the rest of the years, really was a bad time for employment

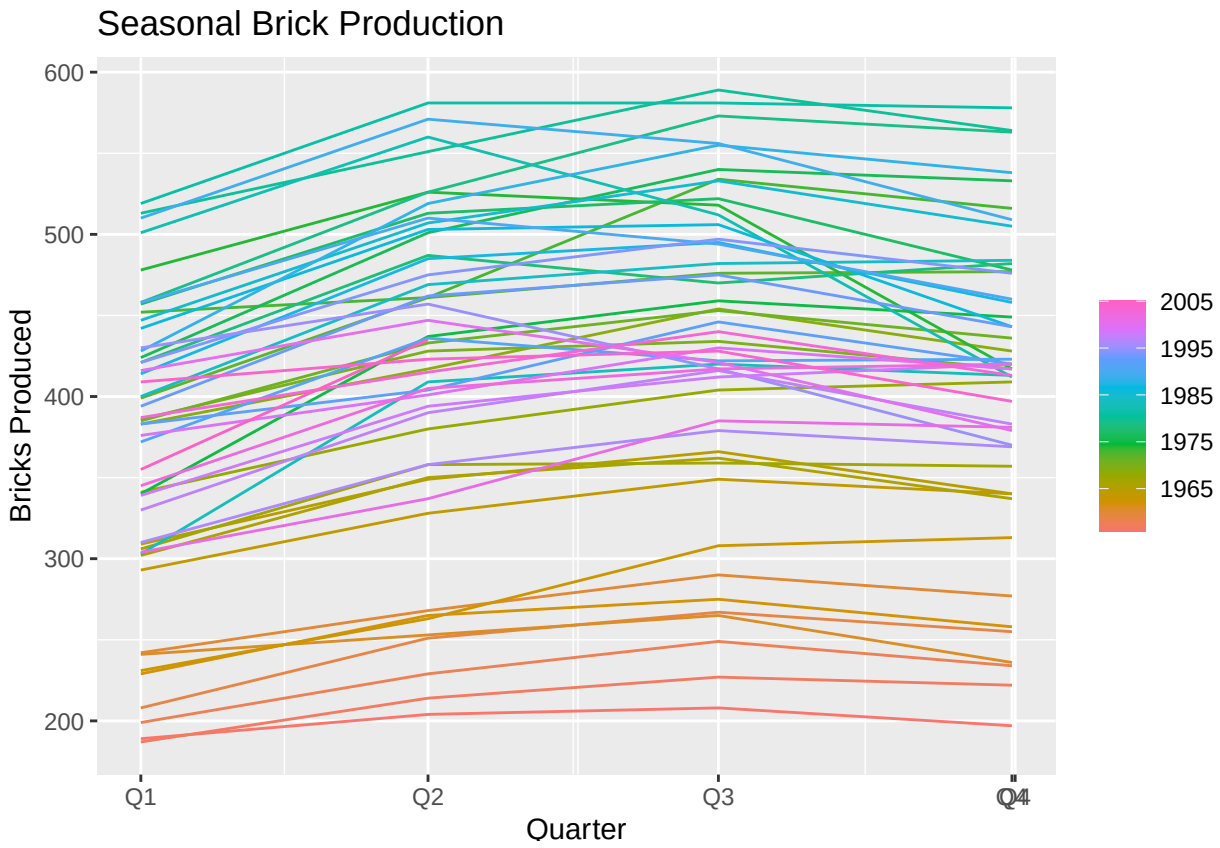
Charting Bricks from aus_production These charts plot Brick production in Australia and covers 1956 Q1 to 2010 Q2

```
aus_production %>%
  drop_na(Bricks) %>%
  autoplot(Bricks) +
  labs(title = "Quarterly Brick Production",
```

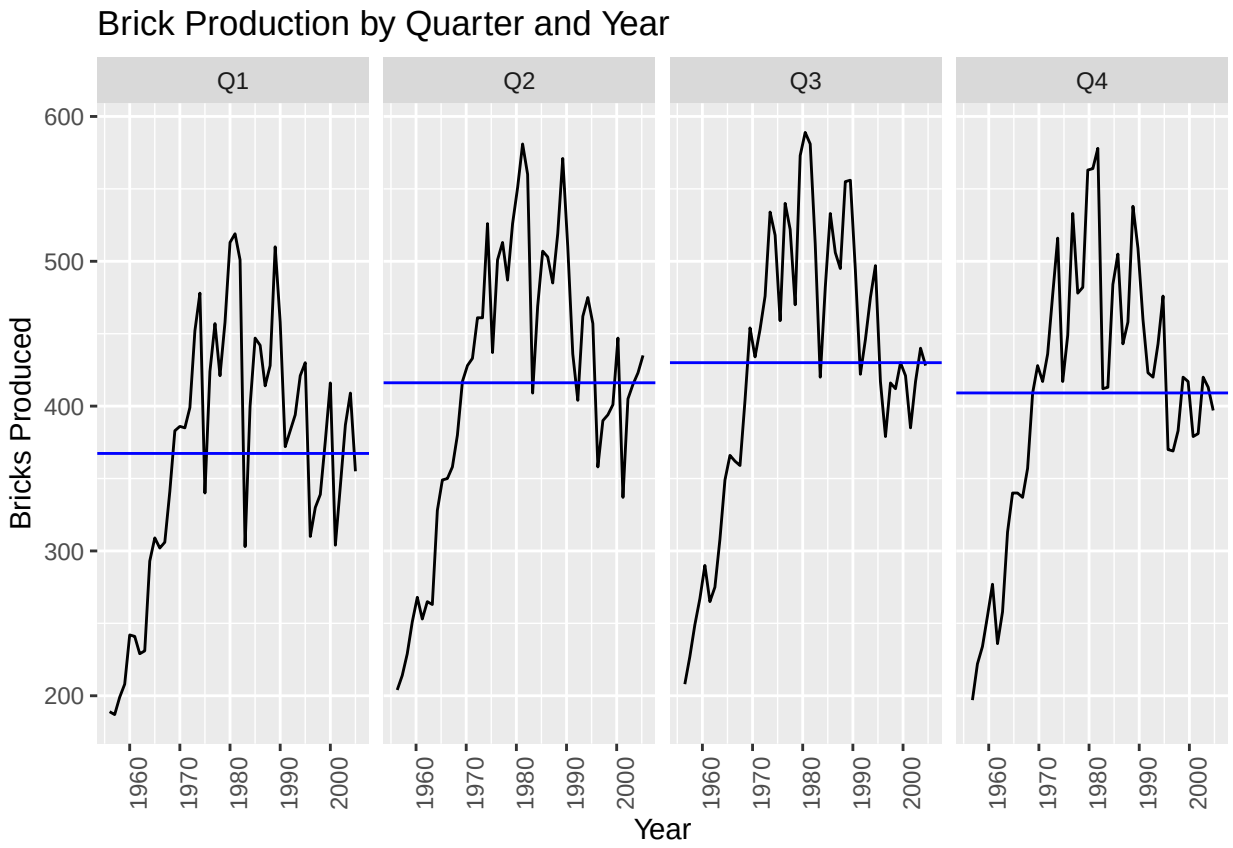
```
  x = "Year & Quarter",
  y = "Bricks Produced")
```



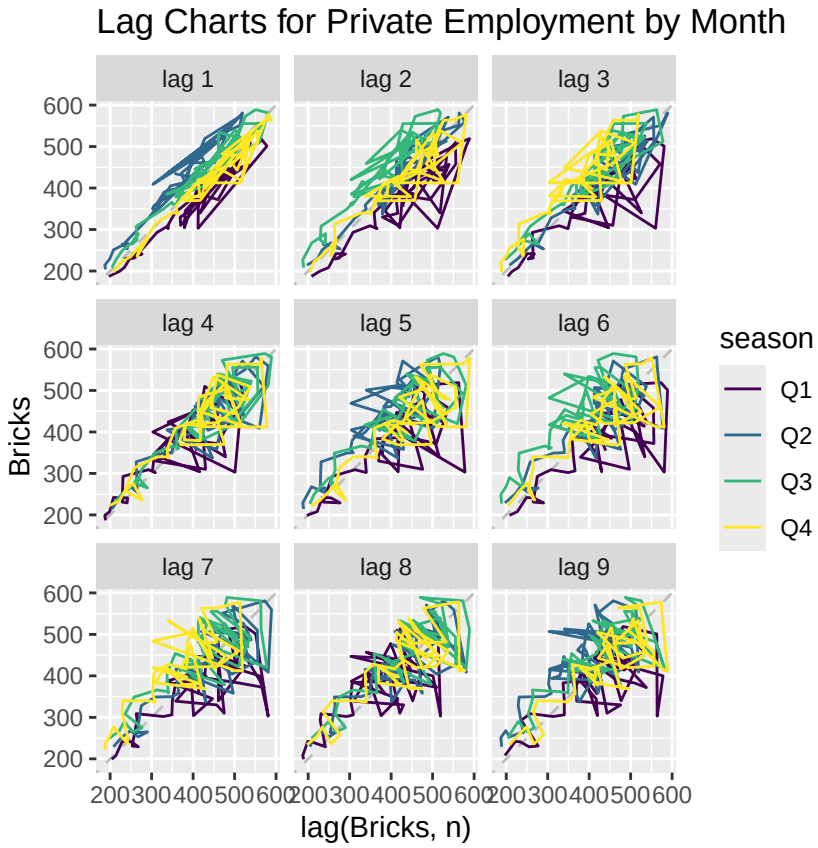

```
aus_production %>%
  drop_na(Bricks) %>%
  gg_season(Bricks) +
  labs(title = "Seasonal Brick Production",
  x = "Quarter",
  y = "Bricks Produced")
```



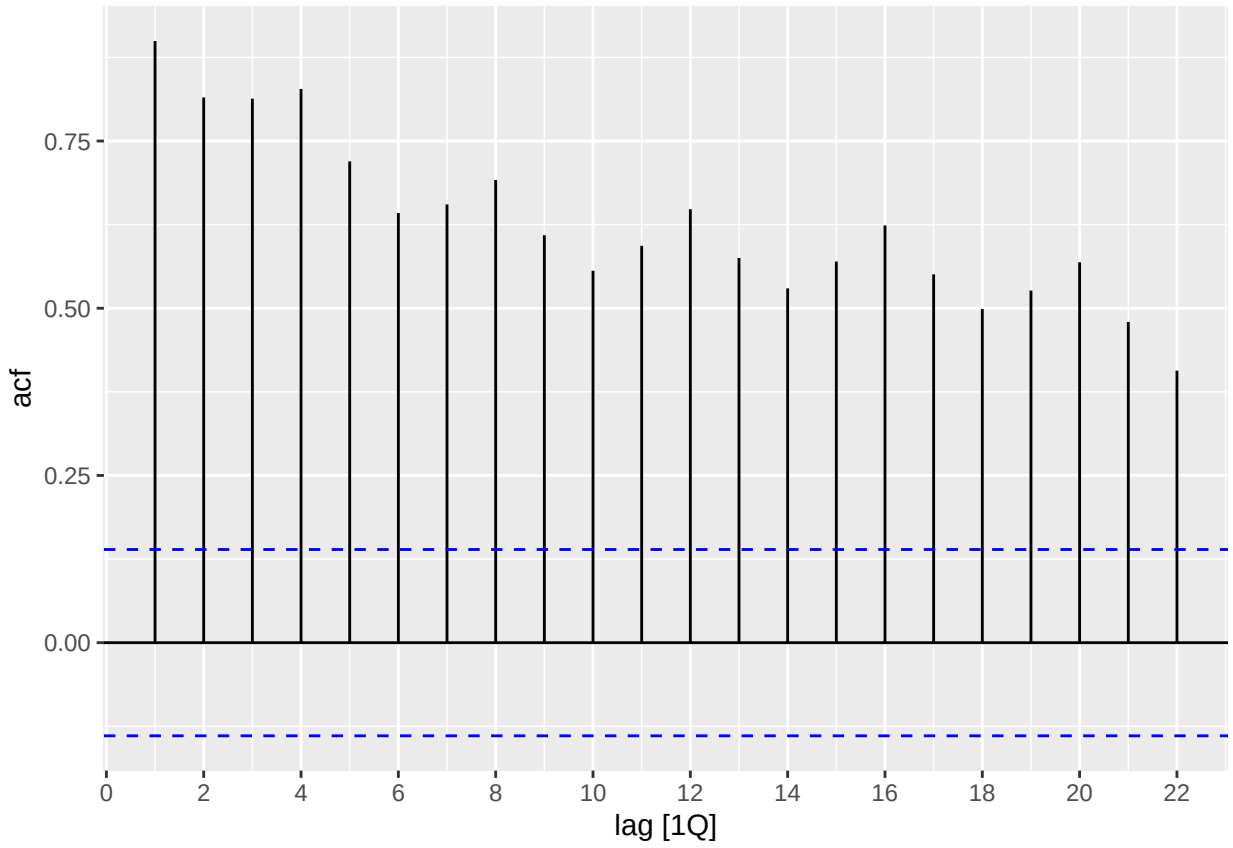

```
aus_production %>%
  drop_na(Bricks) %>%
  gg_subseries(Bricks) +
  labs(title = "Brick Production by Quarter and Year",
  x = "Year",
  y = "Bricks Produced")
```




```
aus_production %>%
  drop_na(Bricks) %>%
  gg_lag(Bricks) +
  labs(title = "Lag Charts for Private Employment by Month")
```




```
aus_production %>%
  drop_na(Bricks) %>%
  ACF(Bricks) %>% autoplot()
```

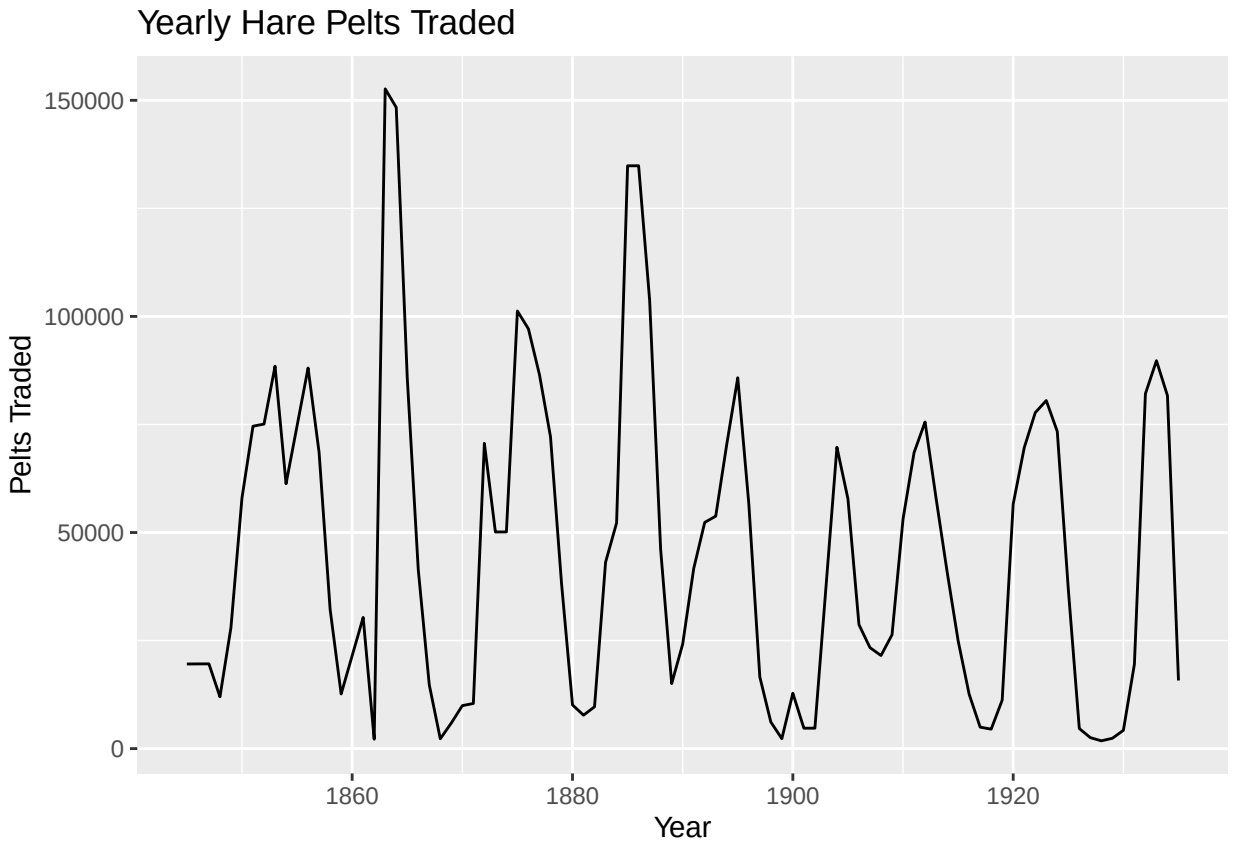


- Can you spot any seasonality, cyclicity and trend?
  - Starting around 1970, quarterly trends bounced around significant but returned to the same 1970 levels by the early 2000's. This suggest some sort of significant change in the 70's and 80's that caused the volume to drop.
  - The lag charts show a decent positive linear trend but there's a lot of noise. The autocorrelation does drop as the lags get bigger. This means that very recent production will influence near term but not long term production.
- What do you learn about the series?
  - Brick production is extremelt volatile and there is most likely some sort of macro pressure that causes the large fluctuations.
  - Forecasting this data would be very challenging. Any given year of production is kind of all over the place and any degree of confidence in a forecast would be on the lower side.
- What can you say about the seasonal patterns?
  - The season patterns are unusual and tied to something economic, rather than the time of year. There's some sort of pressure cause sharp drops and rebounds every few years.
- Can you identify any unusual years?
  - 1983 saw a large drop, followed by an immediate event. This also happen a few years priod ana few years later, suggesting some sort of cyclical pressure on production that plays out over multiple years.

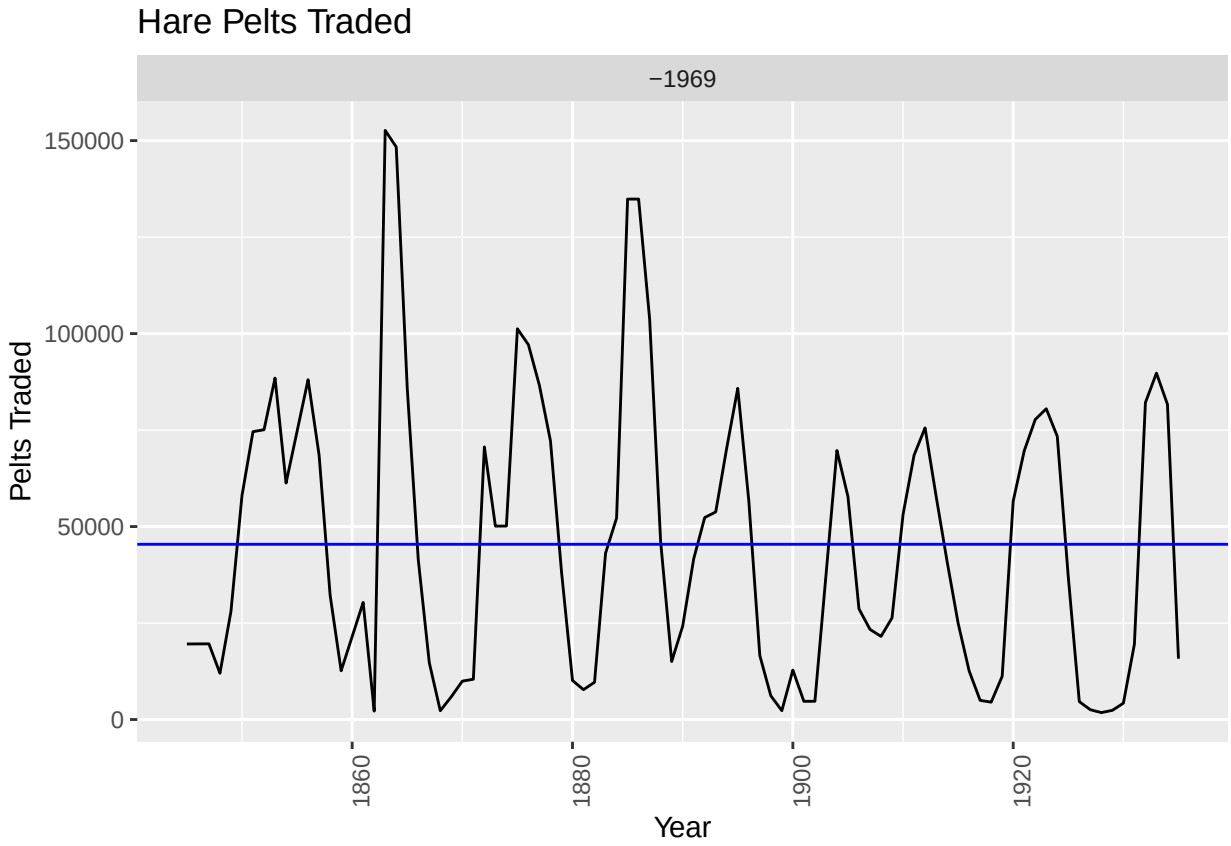
Charting Hare from Pelt Data Set These charts plot the number of Snowshoe Hare pelts traded by the Hudson Bay Company from 1845 to 1935.

```
pelt %>%
  drop_na(Hare) %>%
  autoplot(Hare) +
  labs(title = "Yearly Hare Pelts Traded",
```

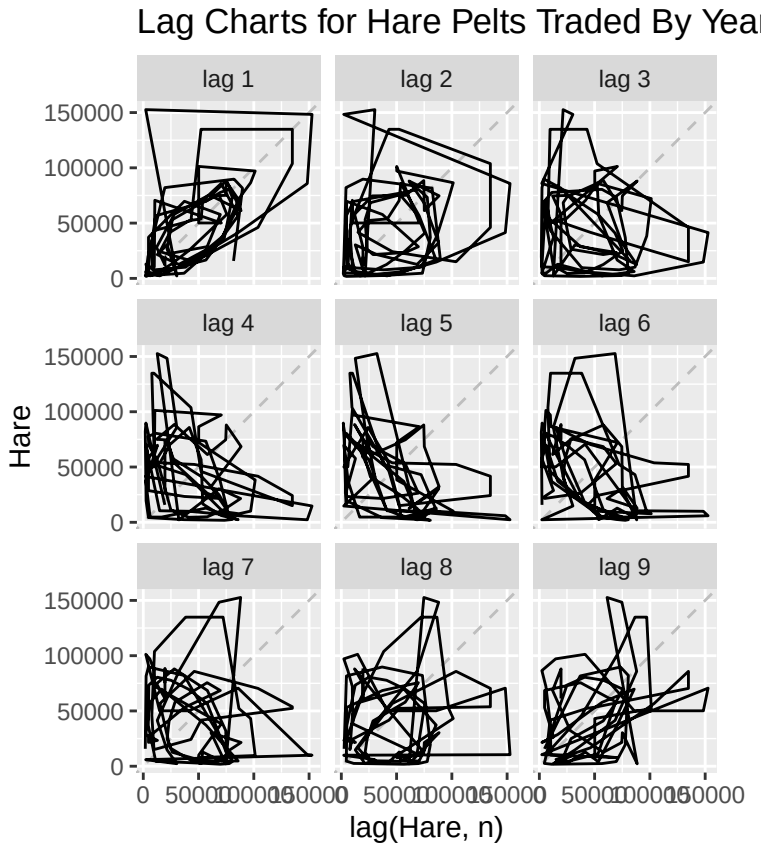
```
  x = "Year",
  y = "Pelts Traded")
```



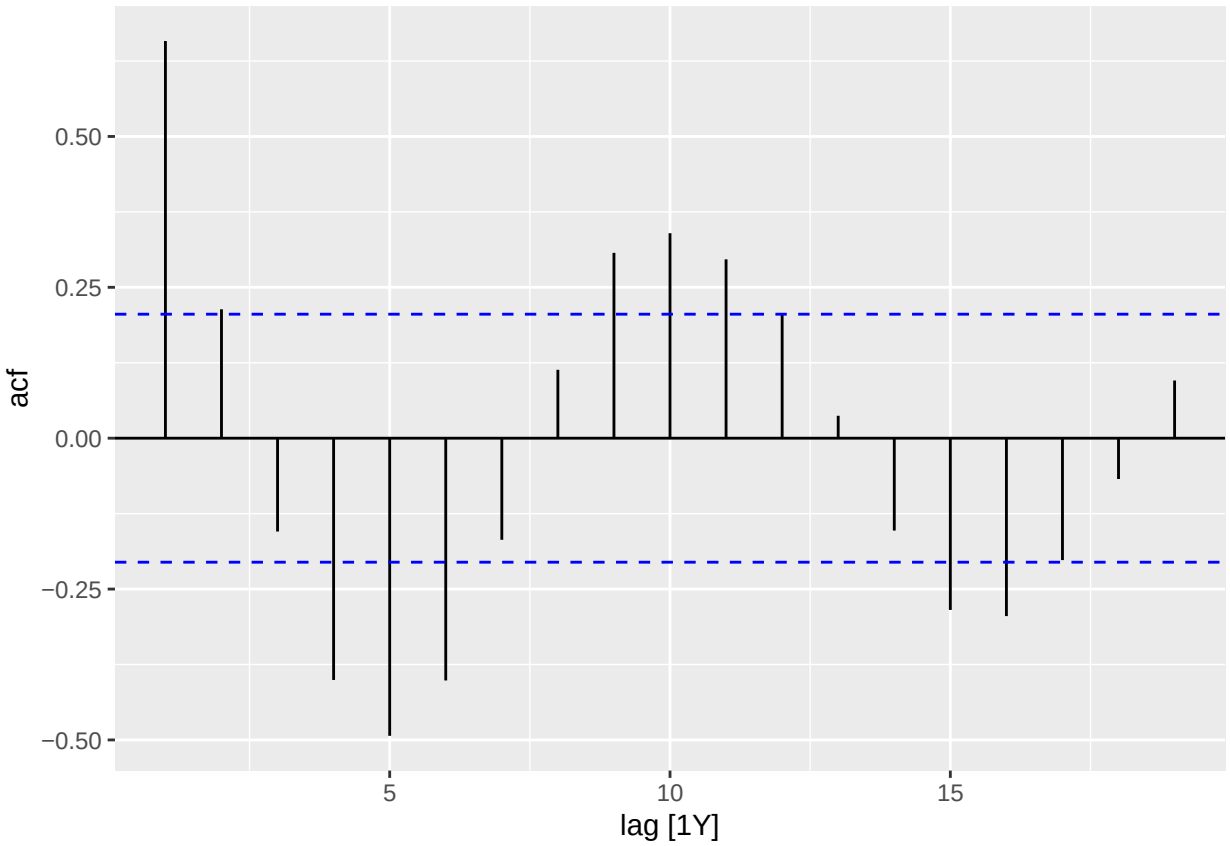

```
pelt %>%
  drop_na(Hare) %>%
  gg_subseries(Hare) +
  labs(title = "Hare Pelts Traded",
  x = "Year",
  y = "Pelts Traded")
```




```
pelt %>%
  drop_na(Hare) %>%
  gg_lag(Hare) +
  labs(title = "Lag Charts for Hare Pelts Traded By Year")
```




```
pelt %>%
  drop_na(Hare) %>%
  ACF(Hare) %>% autoplot()
```

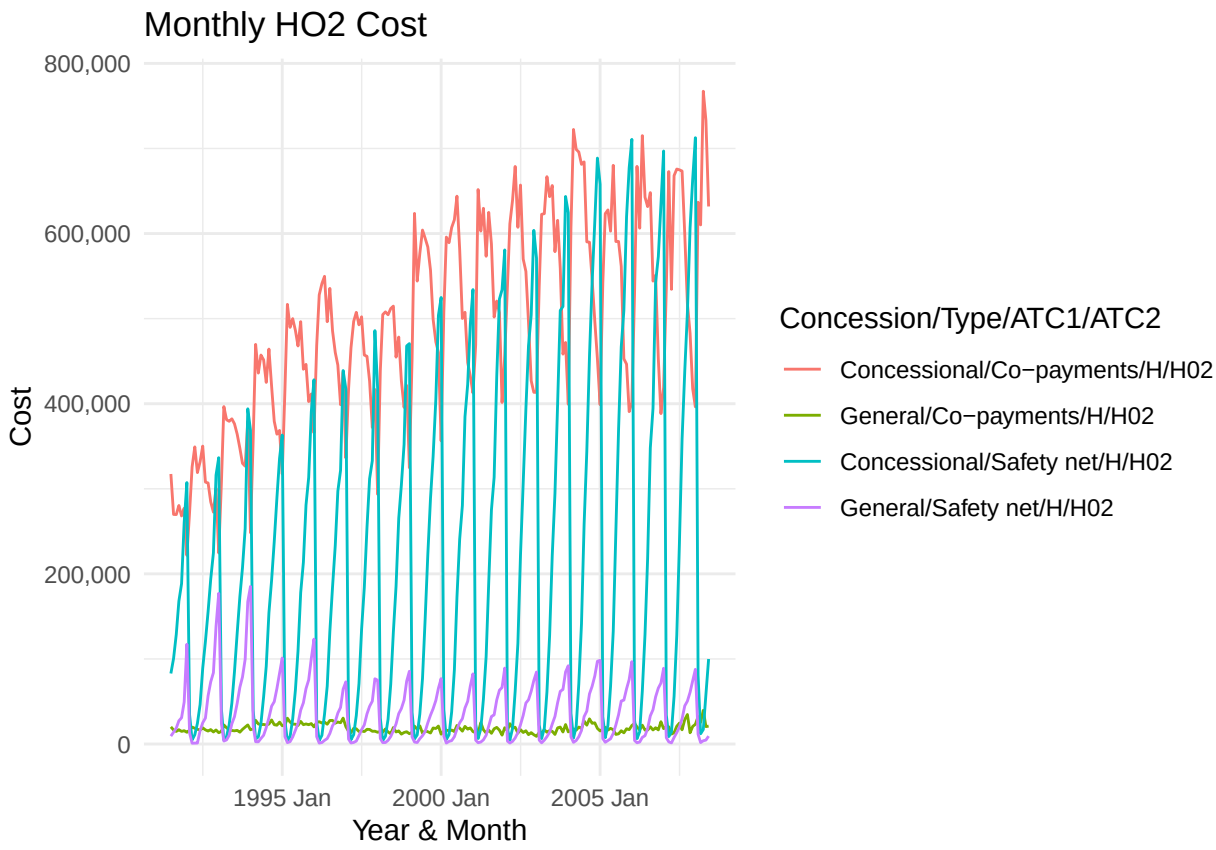


- Can you spot any seasonality, cyclicity and trend?
  - There are sharp peaks and valleys in the data around every ten years
  - The lag data shows that there's basically no relationship between a previous data point and the current one
    - Correlation does change every few years between around -.25 and +.25
  - You could use the multi year cyclical nature
- What do you learn about the series?
  - There are seaonal trends on a multi year cycle that causes sharp spikes and drops
  -
- What can you say about the seasonal patterns?
  - You could use the multi year cyclical nature of the data to sketch out what future cycles would look like
  - There is something deeper that's causing these spikes and valleys. Is there something about Hare pelts that could these types of changes in trading?
- Can you identify any unusual years?
  - I wouldn't say unusual, persay. The cycles happening every few years are interesting and worth understanding more but each year appears to sit where you might expect it to when put in proper context.

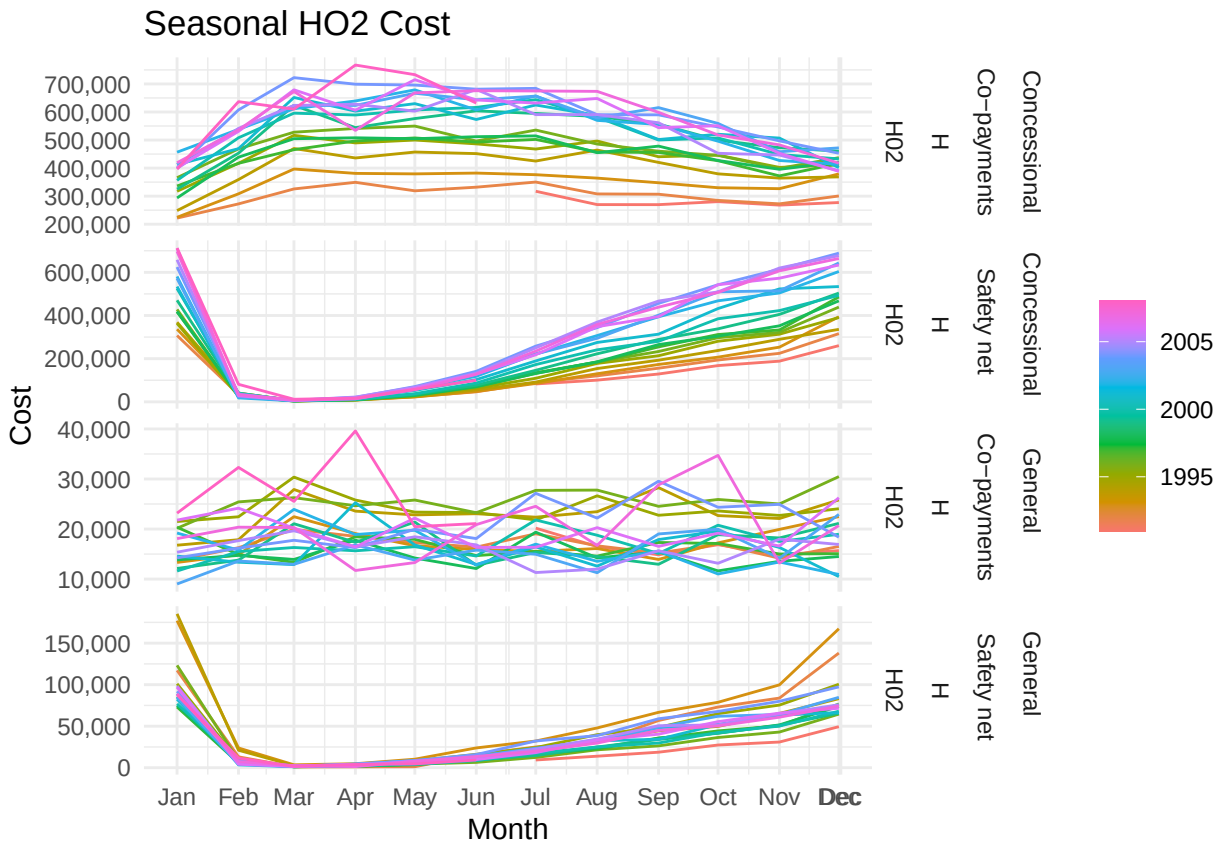
Charting H02 Cost from PBS These charts plot Cost data from the PBS dataset that's been filtered for ATC2 equals "H02".

```
hohtwo <- PBS %>%
  filter(ATC2 == "H02")

hohtwo %>%
  drop_na(Cost) %>%
```

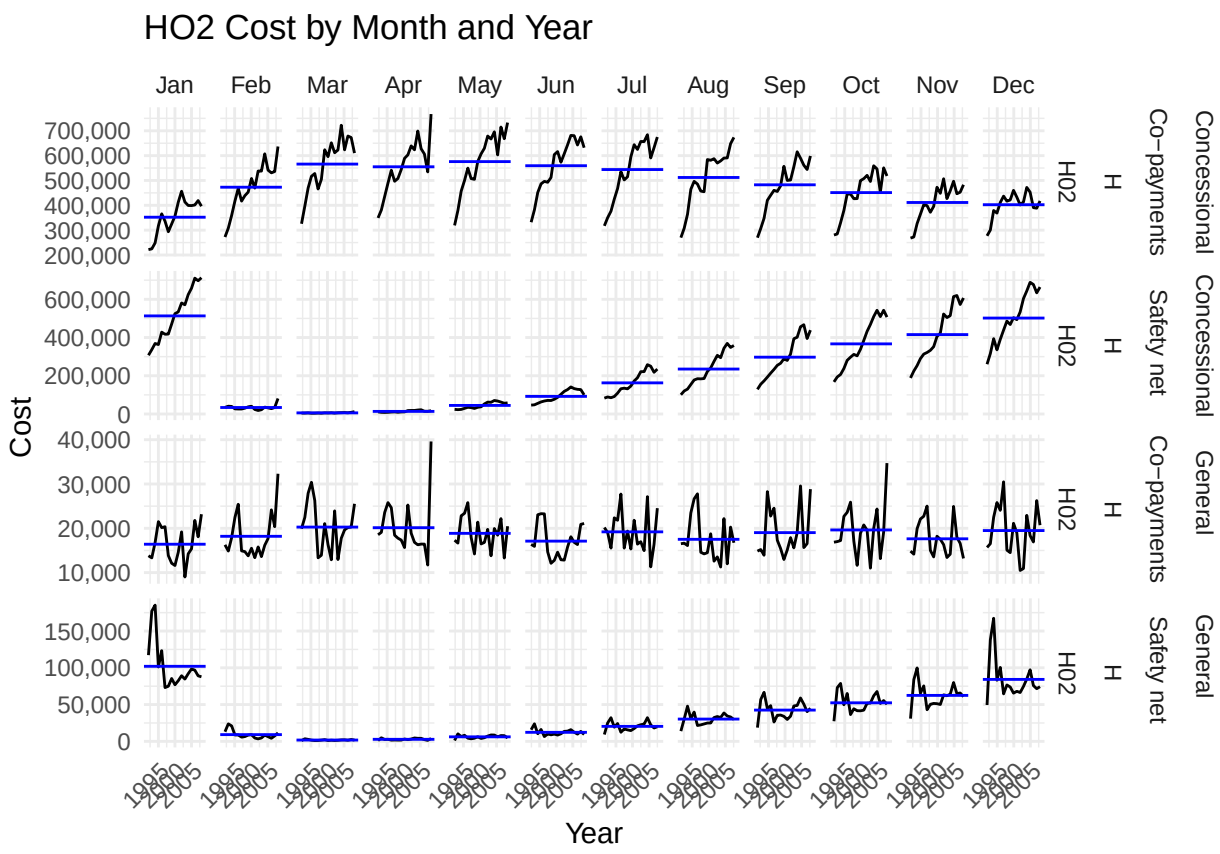
```
  autoplot(Cost) +
  labs(title = "Monthly HO2 Cost",
       x = "Year & Month",
       y = "Cost") +
  scale_y_continuous(labels = scales::label_number(big.mark = ",")) +
  theme_minimal()
```



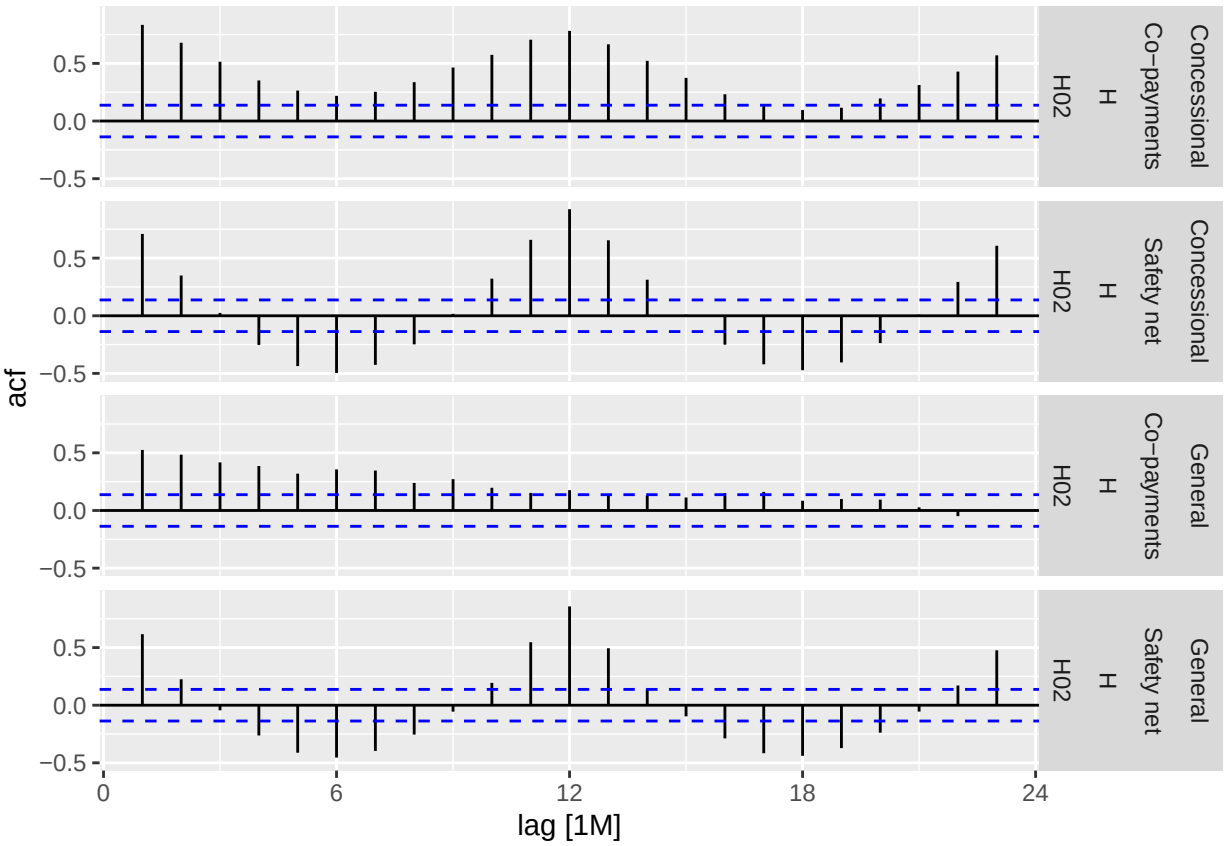

```
hohtwo %>%
  drop_na(Cost) %>%
  gg_season(Cost) +
  labs(title = "Seasonal HO2 Cost",
  x = "Month",
  y = "Cost") +
  scale_y_continuous(labels = scales::label_number(big.mark = ",")) +
  theme_minimal()
```




```
hohtwo %>%
  drop_na(Cost) %>%
  gg_subseries(Cost) +
  labs(title = "HO2 Cost by Month and Year",
       x = "Year",
       y = "Cost") +
  scale_y_continuous(labels = scales::label_number(big.mark = ",")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
hohtwo %>%
  drop_na(Cost) %>%
  ACF(Cost) %>% autoplot()
```

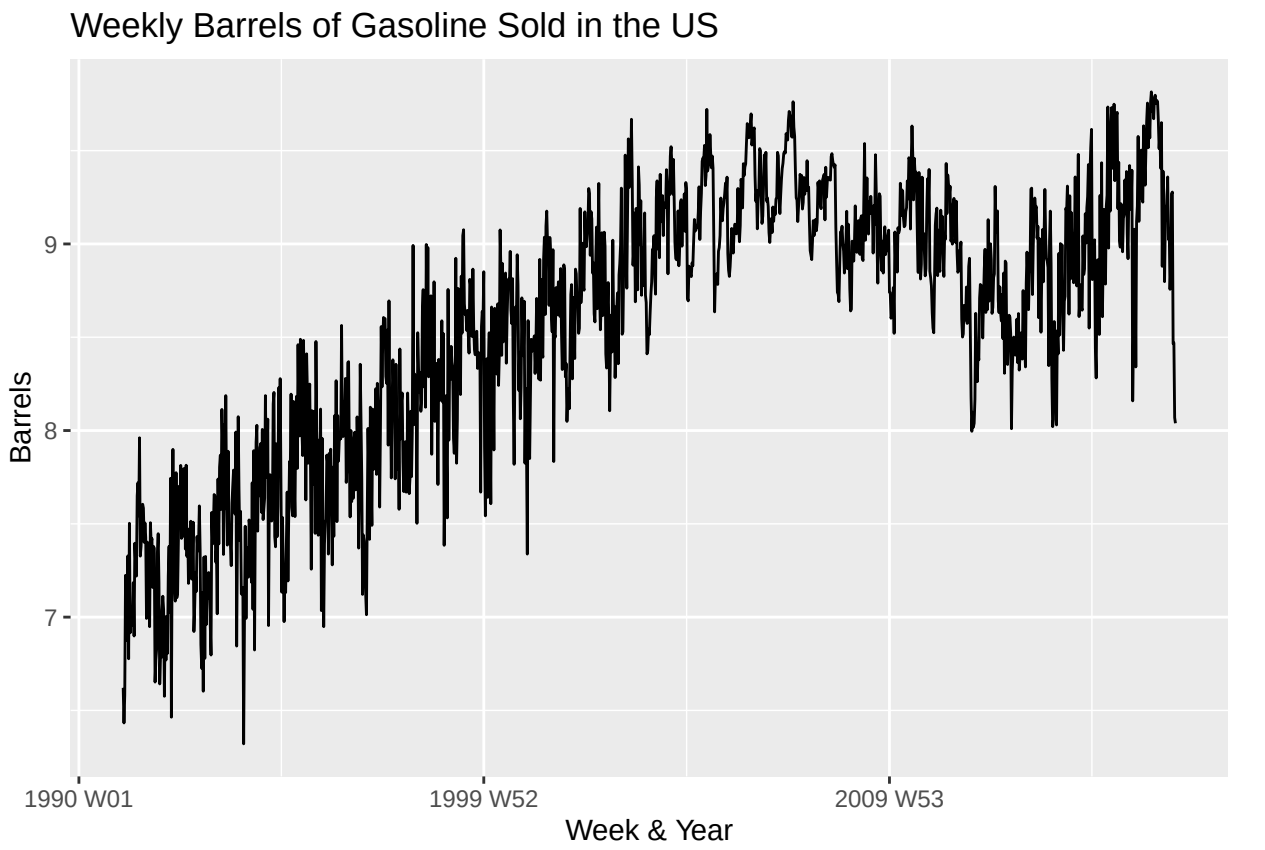


- Can you spot any seasonality, cyclicity and trend? *Safety net for both concessional and general have sharp drops at the start of the year that aren't extensions of drops starting at the end of the year, suggesting some sort of calendar year driver in the drop at the start* The autocorrelations show a cyclical pattern the movement between 0 and .5. Interestingly, the concessional co-payment values always stay positive while the safety net ones go negative.
  - They are moving in roughly the same pattern
- What do you learn about the series? *HO2 costs have been steadily increasing over time, with sharp peaks and valleys on a predictable cycle
- What can you say about the seasonal patterns?
  - There are patterns within a year and also over a longer period of time. You could use this to forecast what the values would be in the future since the pattern has been so consistent
- Can you identify any unusual years?
  - 2005 had something going on that cause weird spikes April and October
- Barrels from us_gasoline

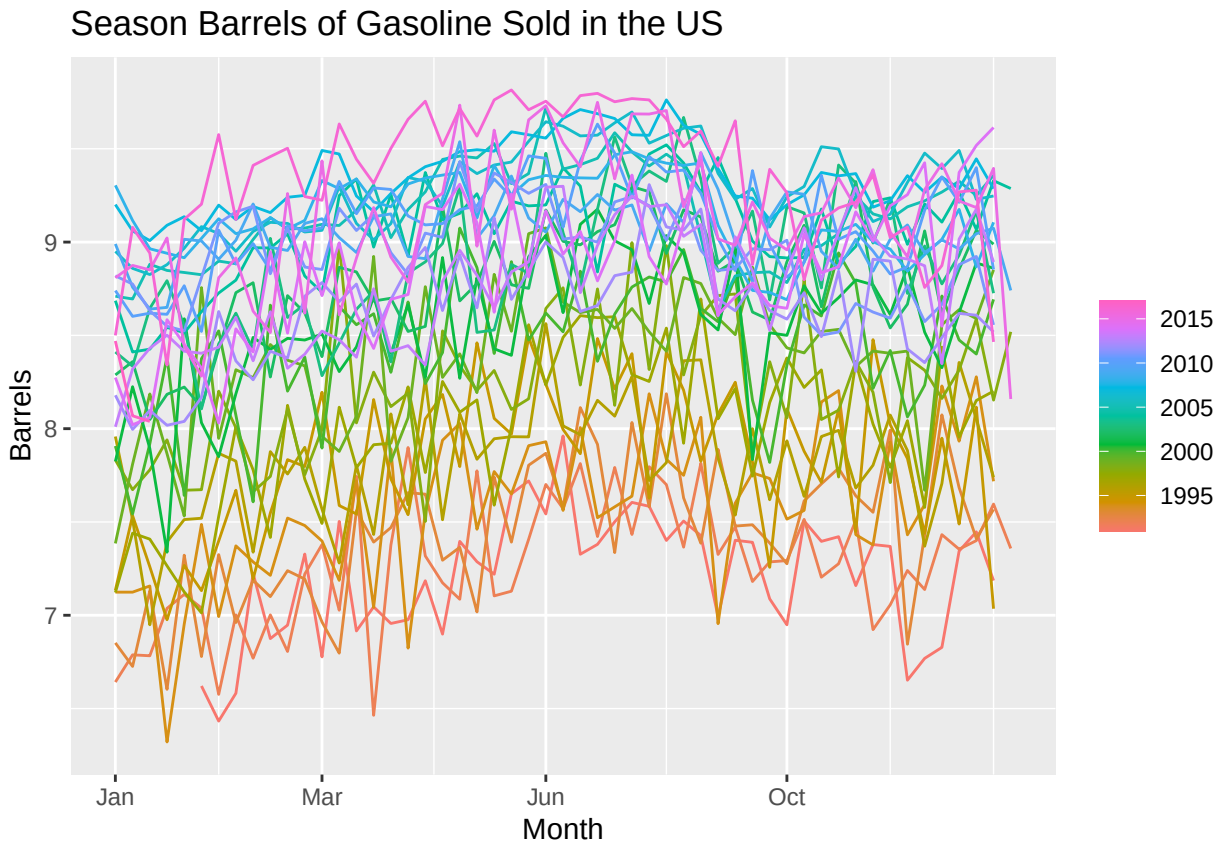
Charting Barrels fron us_gasoline These are charts showing millions of barrels of gasoline sold per week in the United STates between Week 6 1991 and Week 3 2017. All Barrels sold are in units of millions of barrels.

```
us_gasoline %>%
  drop_na(Barrels) %>%
  autoplot(Barrels) +
  labs(title = "Weekly Barrels of Gasoline Sold in the US",
```

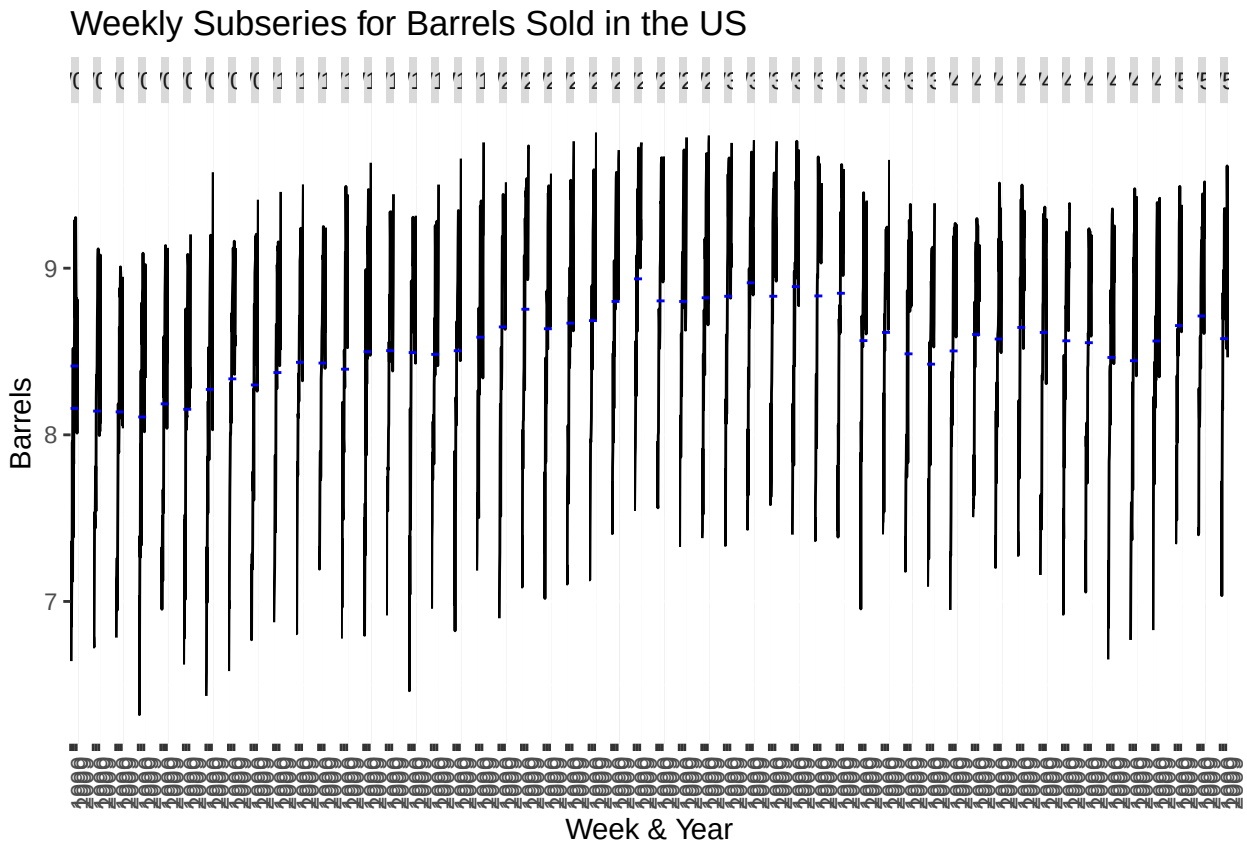
```
  x = "Week & Year",
  y = "Barrels")
```



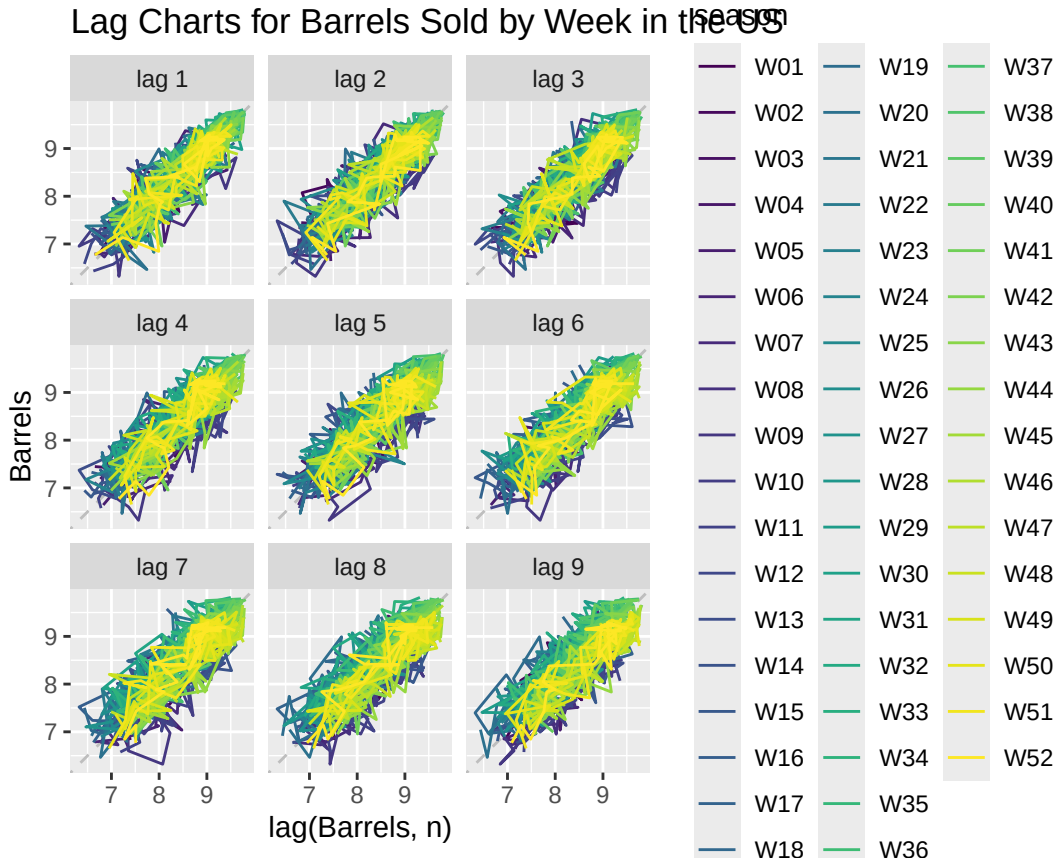

```
us_gasoline %>%
  drop_na(Barrels) %>%
  gg_season(Barrels) +
  labs(title = "Season Barrels of Gasoline Sold in the US",
  x = "Month",
  y = "Barrels")
```



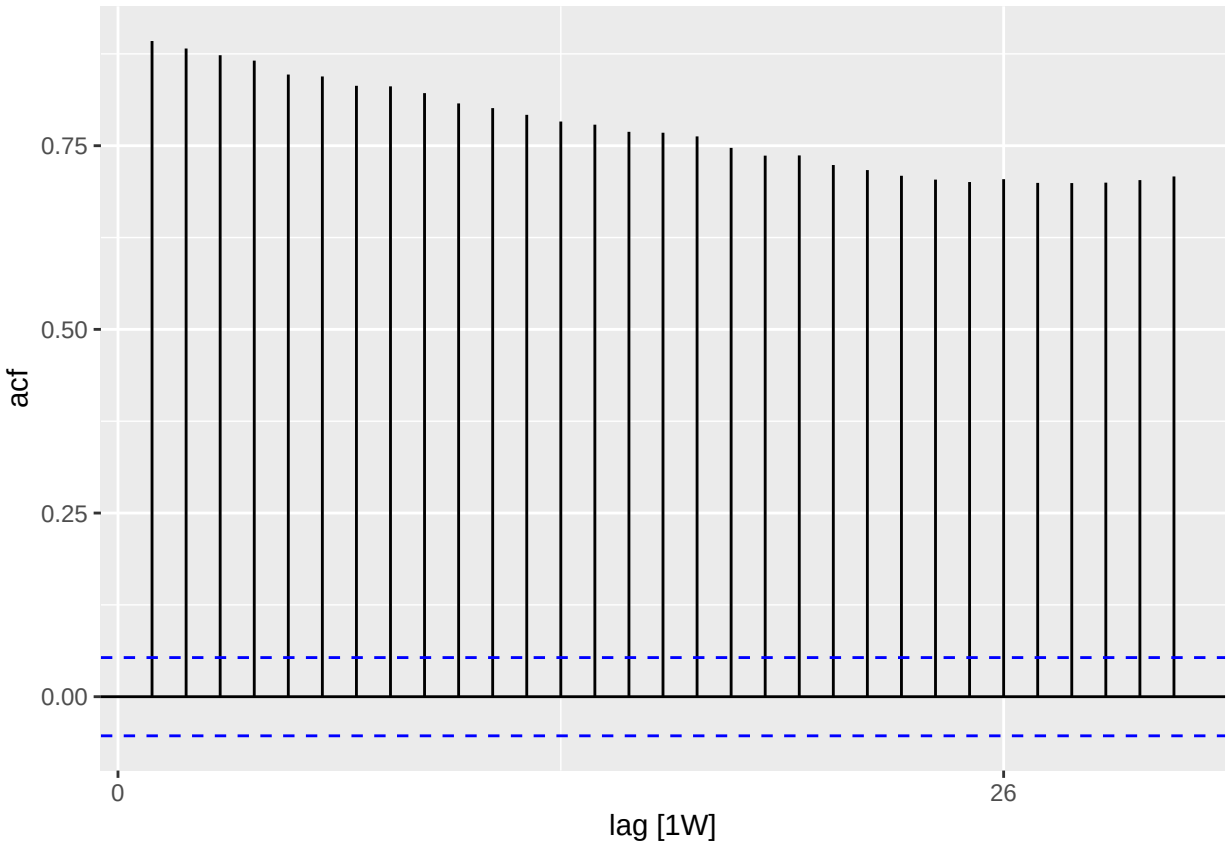

```
us_gasoline %>%
  drop_na(Barrels) %>%
  gg_subseries(Barrels) +
  labs(title = "Weekly Subseries for Barrels Sold in the US",
  x = "Week & Year",
  y = "Barrels")
```




```
us_gasoline %>%
  drop_na(Barrels) %>%
  gg_lag(Barrels) +
  labs(title = "Lag Charts for Barrels Sold by Week in the US")
```




```
us_gasoline %>%
  drop_na(Barrels) %>%
  ACF(Barrels) %>% autoplot()
```



- Can you spot any seasonality, cyclicity and trend? *There's a strong positive linear relationship for all lag values and for most weeks, with corresponsing high correlation values
  - Interestingly, the correlation values trend down as you progress through the year
  - There are positive trends in the US summer months, which is something we hear about in the press.
- What do you learn about the series?
  - Americans do love gasoline, with weekly sales volume incerasing from 6.6 million in the early 1991 to 8 million in 2017 *Sales by week is a strong way to figue out what values will be in the same week in a future year
- What can you say about the seasonal patterns?
  - The consistent drops around the holiday break in the US that's matched with a sharp spike at the start of the year shows that driving patterns are pretty predictable through basic common sense.
  - It's puzzling that the major recession years, like 2008-2010, didn't seem to cause much of a change
- Can you identify any unusual years?
  - 2005-2009 because something happened to cause a general flat lining of sales over a 10-15 year stretch
    - The consistent upwards trend seen before the mid 2000s stopped