624_wk3_E_Haley

Ethan Haley

2/7/2022

```
library(fpp3)
```

```
## -- Attaching packages --------------------------------------------- fpp3 0.4.0 --

## v tibble      3.1.5     v tsibble     1.1.1
## v dplyr       1.0.7     v tsibbledata 0.4.0
## v tidyr       1.1.4     v feasts      0.2.2
## v lubridate   1.8.0     v fable       0.3.1
## v ggplot2     3.3.5

## Warning: package 'tsibbledata' was built under R version 4.0.5

## -- 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()
```

1) Consider the GDP information in `global_economy`. Plot the GDP per capita for each country over time. Which country has the highest GDP per capita? How has this changed over time?

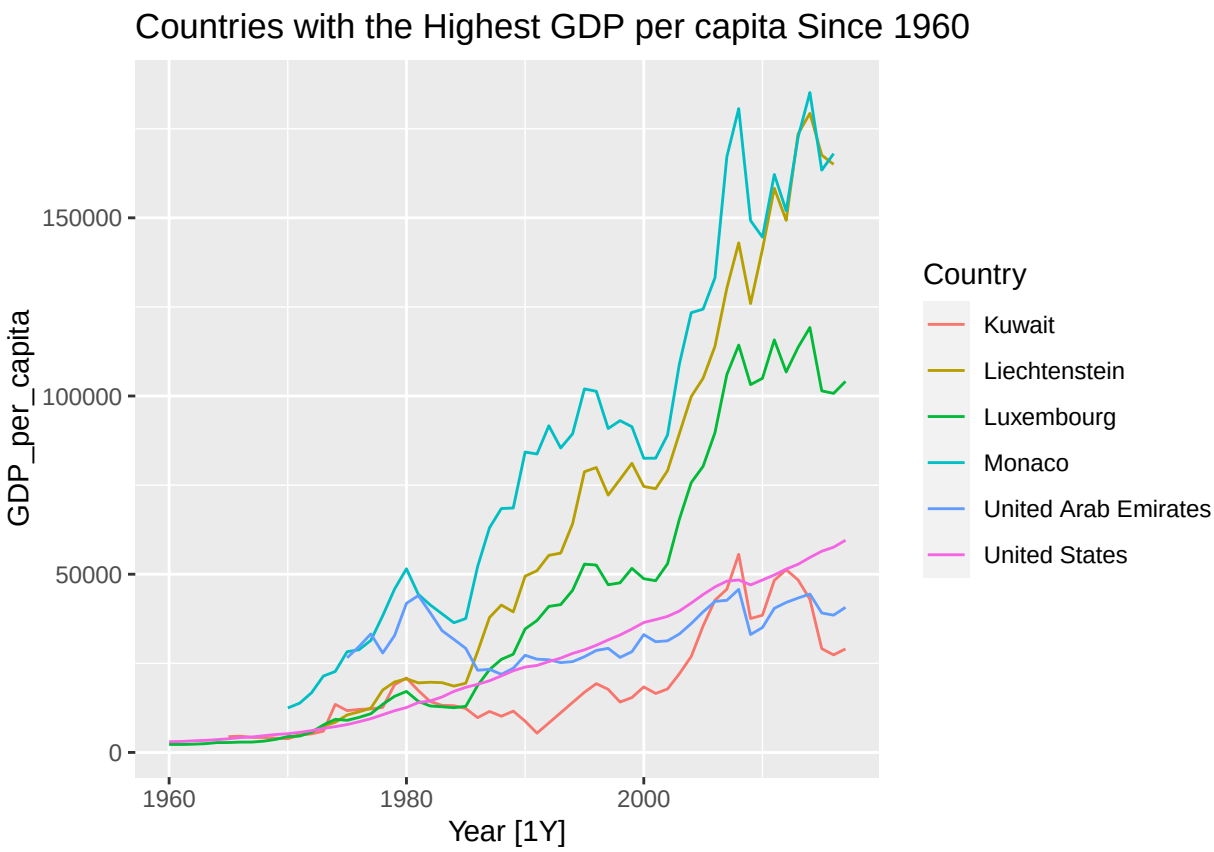
Since there are too many countries to make sense of on one plot, I'll compare all countries that have ever had the highest GDP per capita in any one year.

```
# Calculate the desired metric
global_economy %>%
  mutate(GDP_per_capita = GDP / Population) -> myGDP
# Find out which countries have been the highest,
## since there are too many to plot together
myGDP %>%
  index_by(Year) %>%
  slice_max(GDP_per_capita) -> maxGDPpc

maxGDPs = unique(maxGDPpc$Country)

myGDP %>%
  filter(Country %in% maxGDPs) %>%
```

```
  filter(!is.na(GDP_per_capita)) %>%
  autoplot(GDP_per_capita) +
  labs(title = "Countries with the Highest GDP per capita Since 1960")
```



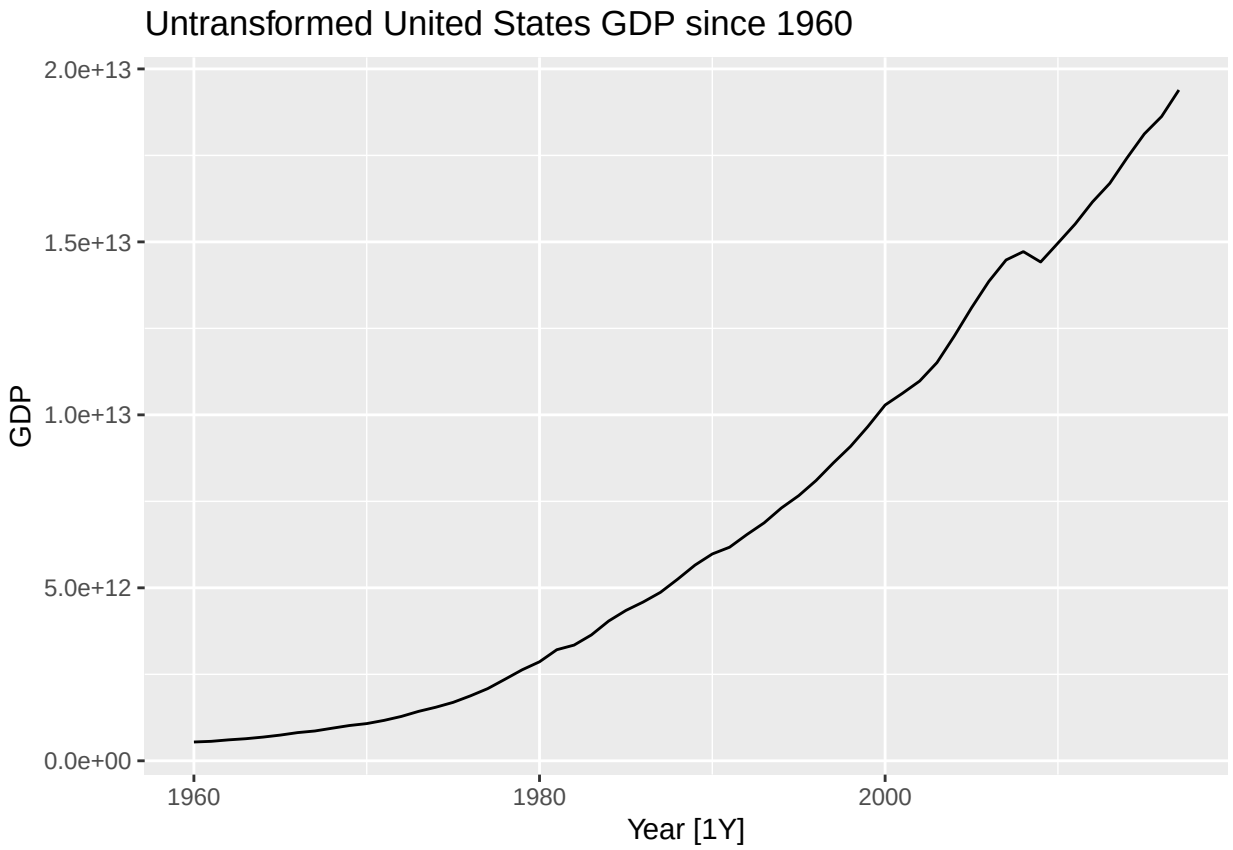
Monaco has always been near the top. The other 2 small, wealthy European countries have remained relatively close to it, with Liechtenstein even catching up to it in the most recent years of data. The 2 wealthy Gulf oil countries have mostly trended more slowly upward in a second trio/tier of countries that includes the U.S., which has the most steady numbers, from year to year.

2) For each of the following series, make a graph of the data. If transforming seems appropriate, do so and describe the effect.

United States GDP from `global_economy`

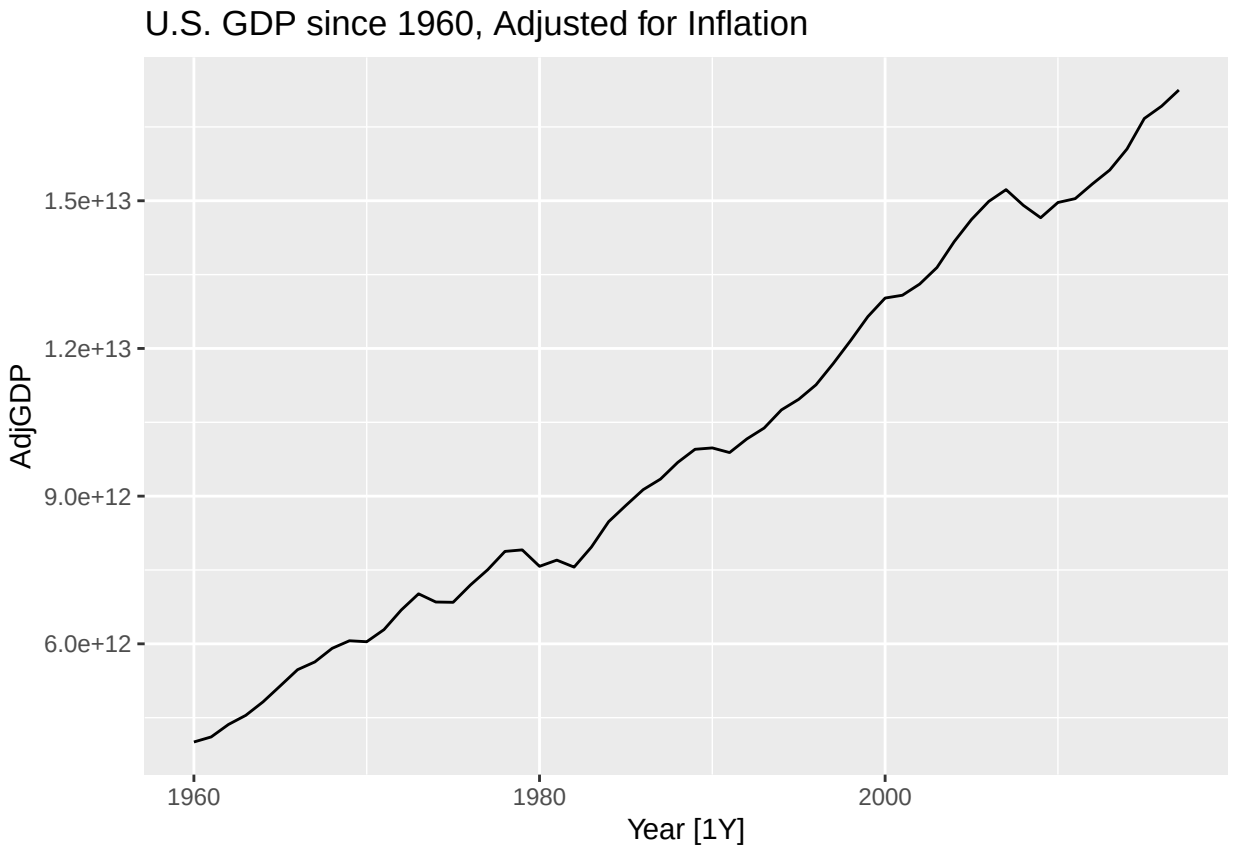
Untransformed:

```
global_economy %>%
  filter(Country=="United States") %>%
  autoplot(GDP)+
  labs(title = "Untransformed United States GDP since 1960")
```



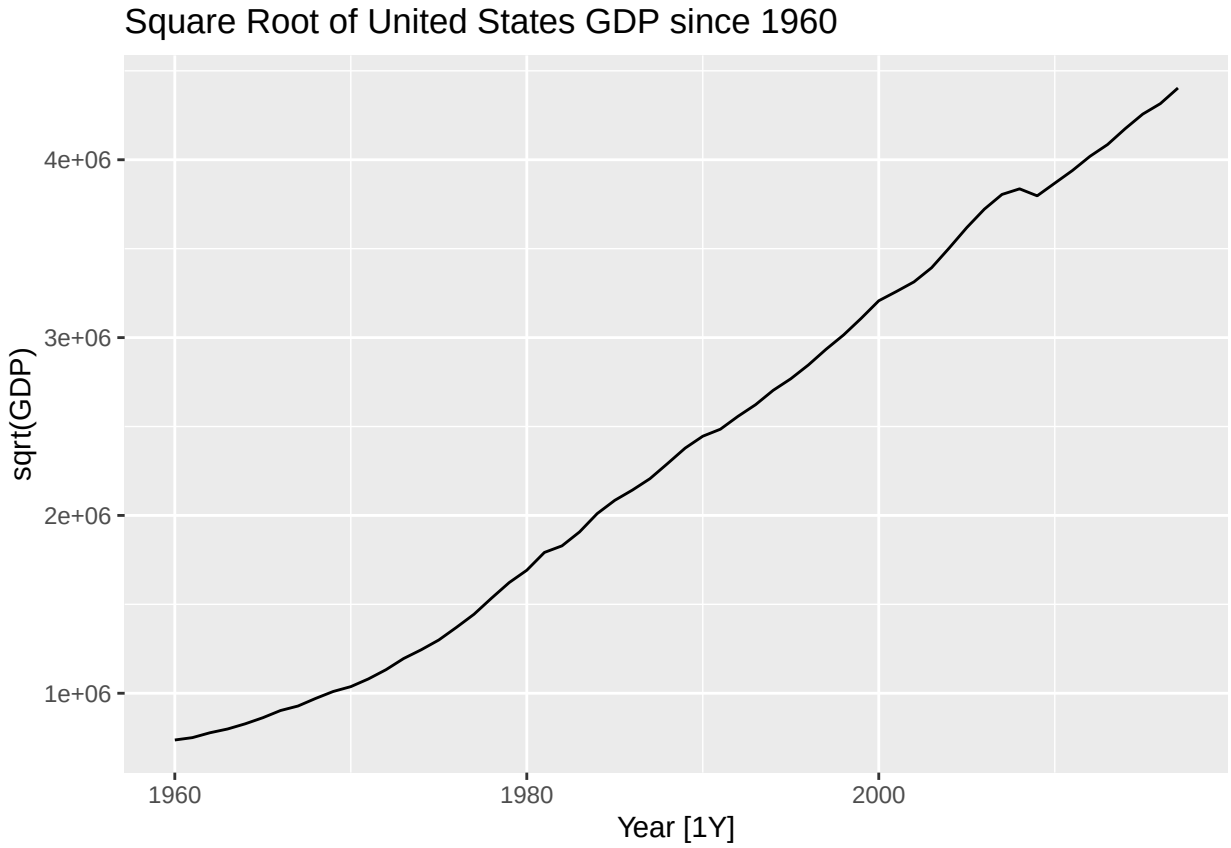
How does that look when adjusted for inflation?

```
global_economy %>%
  filter(Country=="United States") %>%
  mutate(AdjGDP = GDP / CPI * 100) %>%
  autoplot(AdjGDP) +
  labs(title = "U.S. GDP since 1960, Adjusted for Inflation")
```



That looks much more linear, so might work better in certain models. But for an even smoother line, a power transformation might work. Here's how the square root of the GDP looks, without adjusting for inflation:

```
global_economy %>%
  filter(Country=="United States") %>%
  autoplot(sqrt(GDP))+
  labs(title = "Square Root of United States GDP since 1960")
```

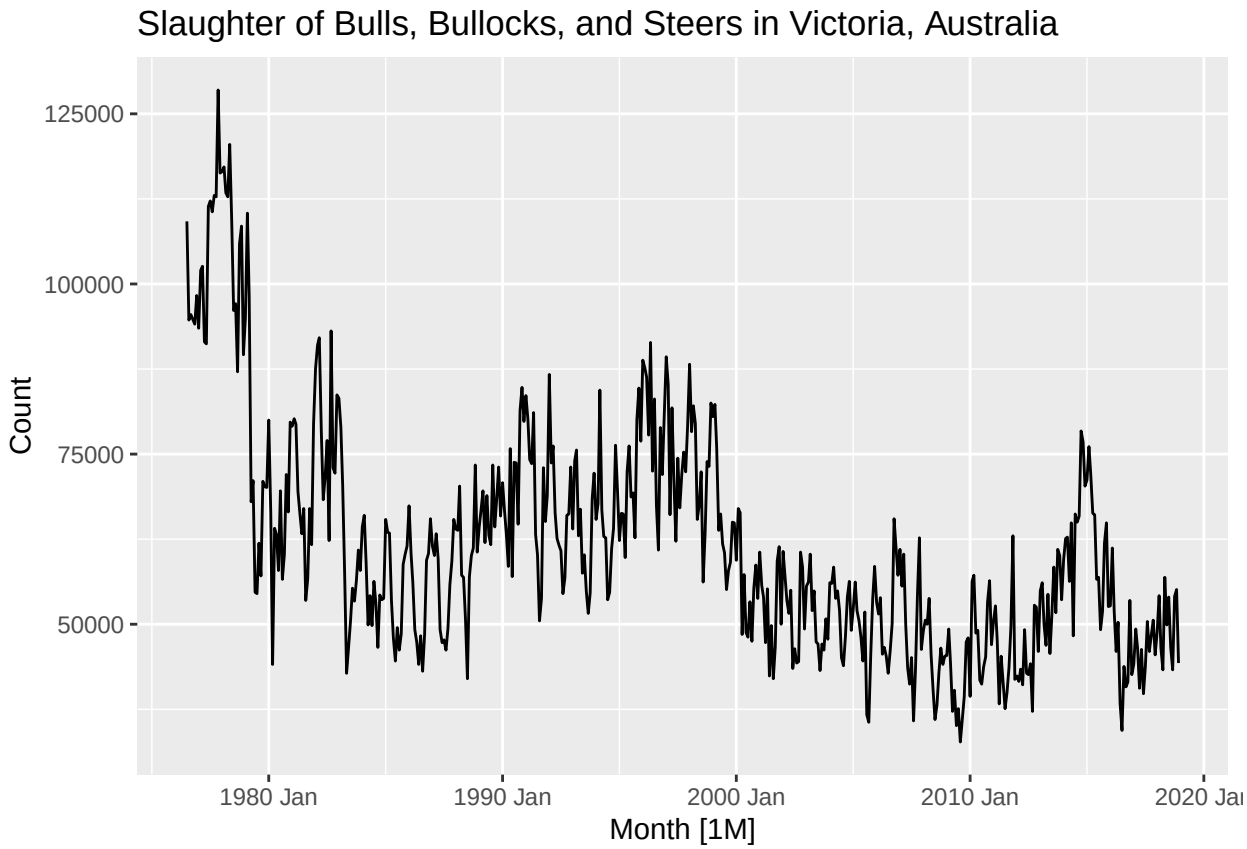


There's not necessarily any inherent reason to transform the GDP numbers by using their square roots, but this line is undoubtedly smoother since 1970, when compared to the inflation-adjusted one. Without knowing much about the economy, I'd suspect that the relative jaggedness of the inflation-adjusted graph is due to the CPI numbers being inconsistent.

Slaughter of Victorian "Bulls, bullocks and steers" in `aus_livestock`.

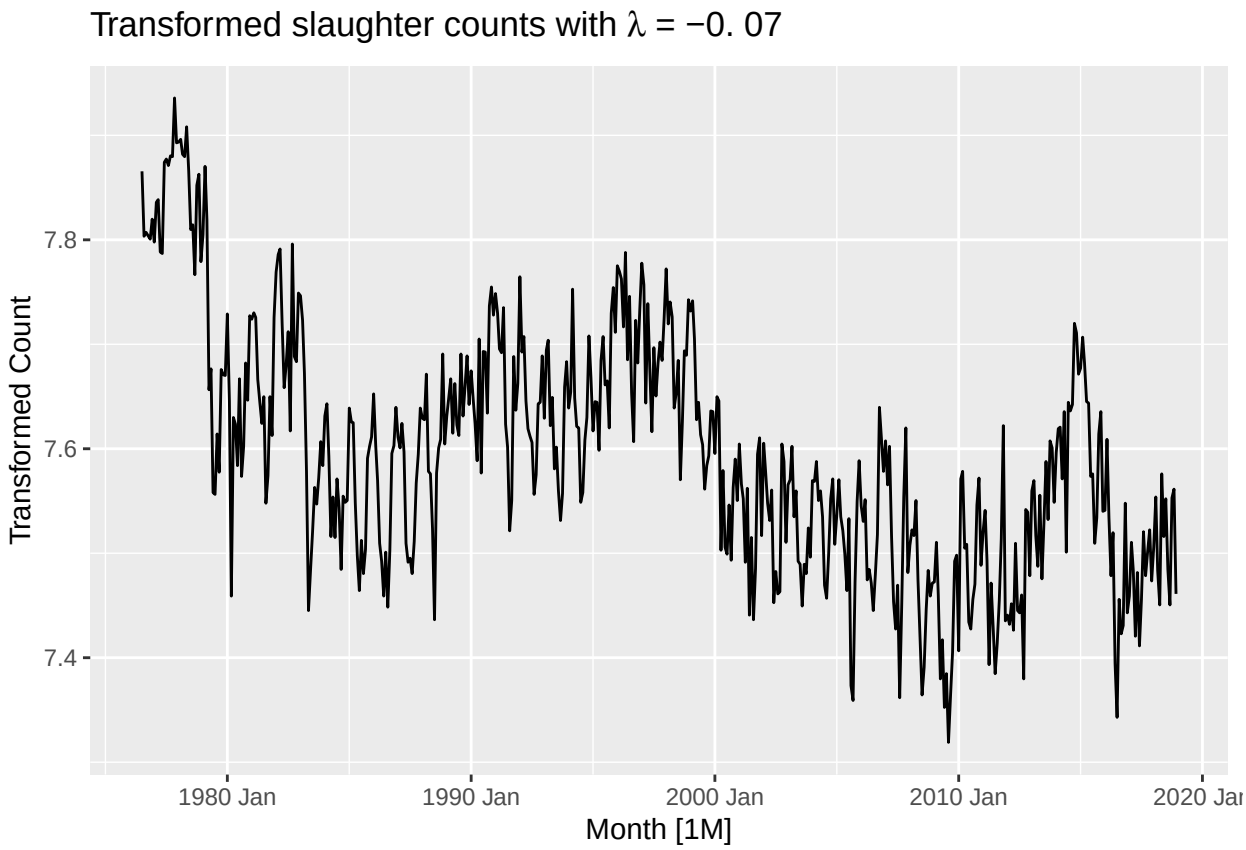
Untransformed data:

```
vicBull = aus_livestock %>%
  filter(Animal=="Bulls, bullocks and steers") %>%
  filter(State=='Victoria')
vicBull %>%
  autoplot(Count) +
  labs(title = "Slaughter of Bulls, Bullocks, and Steers in Victoria, Australia")
```



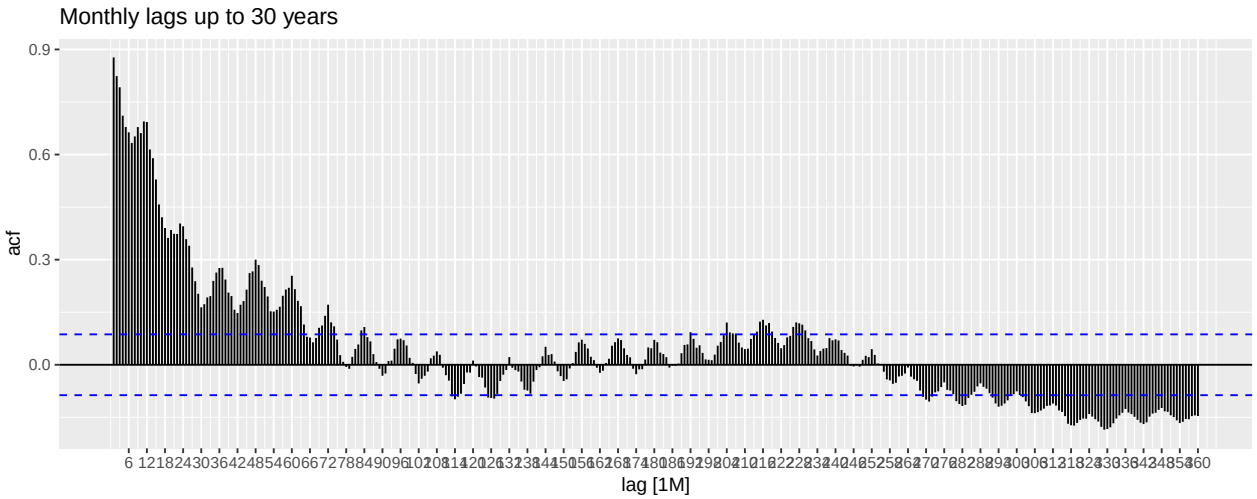
The seasonal swings shown above have generally decreased from decade to decade, such that the Guerrero Method code shown in section 3.1 of our book may help
even things out via a Box-Cox power transformation:

```
# https://otexts.com/fpp3/transformations.html
lambda <- vicBull %>%
  features(Count, features = guerrero) %>%
  pull(lambda_guerrero)
vicBull %>%
  autoplot(box_cox(Count, lambda)) +
  labs(y = "Transformed Count",
       title = latex2exp::TeX(paste0(
         "Transformed slaughter counts with $\\lambda$ = ",
         round(lambda,2))))
```



Another visible feature of the data is the approximately 20 year cycle of increasing and decreasing numbers. Let's see how much the lags are correlated:

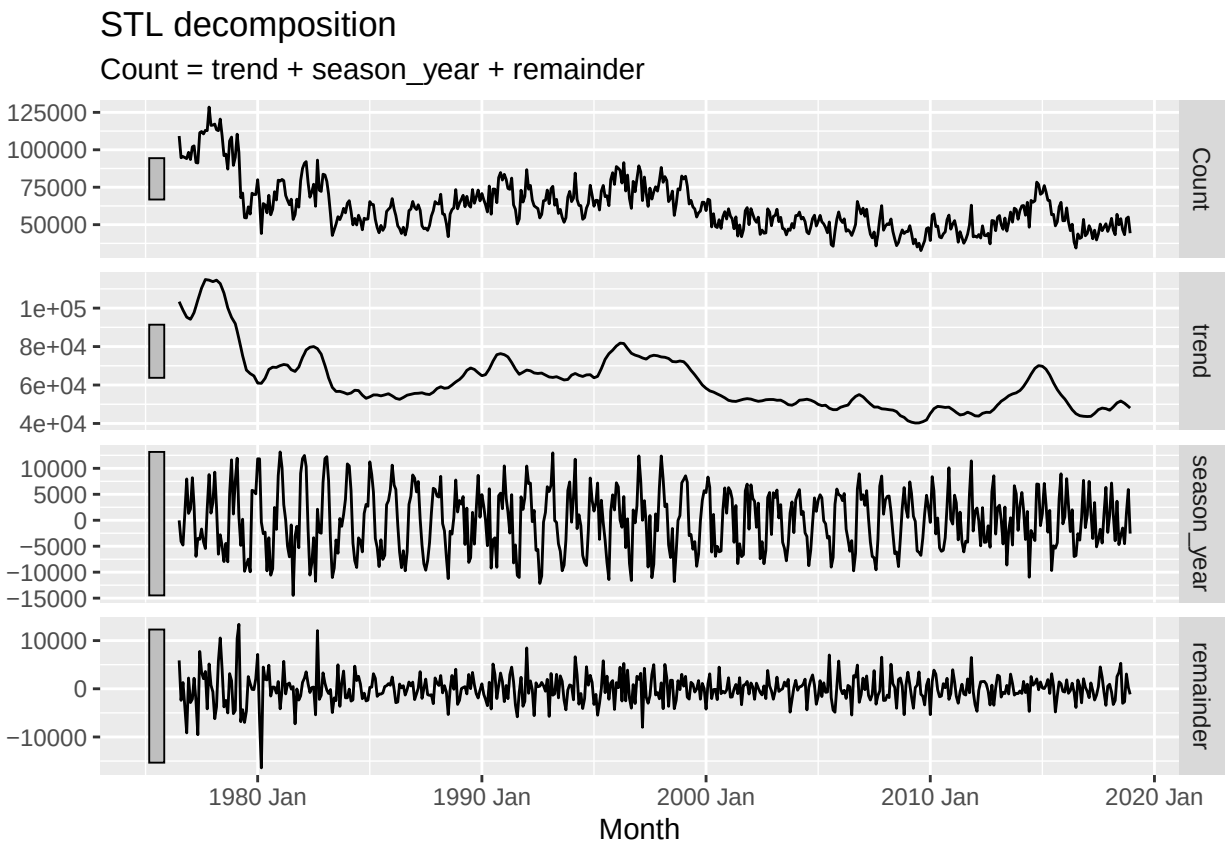
```
vicBull %>%
  ACF(Count, lag_max =360) %>%
  autoplot() + labs(title = "Monthly lags up to 30 years")
```



The highest autocorrelation beyond 5 years is actually the negative correlation between datapoints separated by 25-30 years. There are only about 42 years of data here, so the number of such lags is somewhat limited.

Finally, just to see how the data might break down into seasonal, trend, and remainder components:

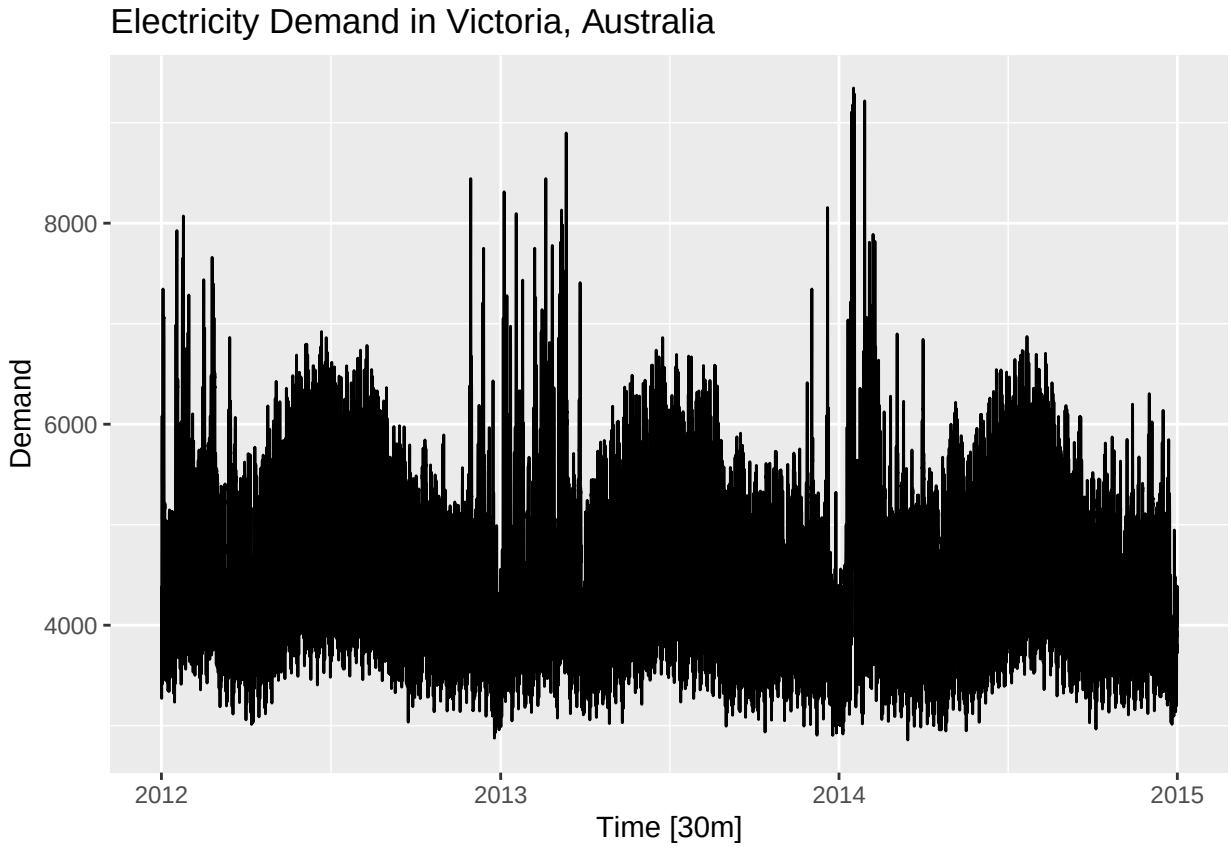
```
dcmp <- vicBull %>%
  model(STL(Count ~ trend(window=11) + season(window=5)))
components(dcmp) %>%
  autoplot()
```



The large remainders near the left of the graph result from wild yearly swings in the numbers of bulls slaughtered in that era. Specifying the seasonal window to be 5 years, instead of the default of 13 years, allows the seasonal component to vary more, and reduces the extremeness of the remainder components (they are more extreme than shown, with the default settings of `STL`).

Victorian Electricity Demand from `vic_elec`

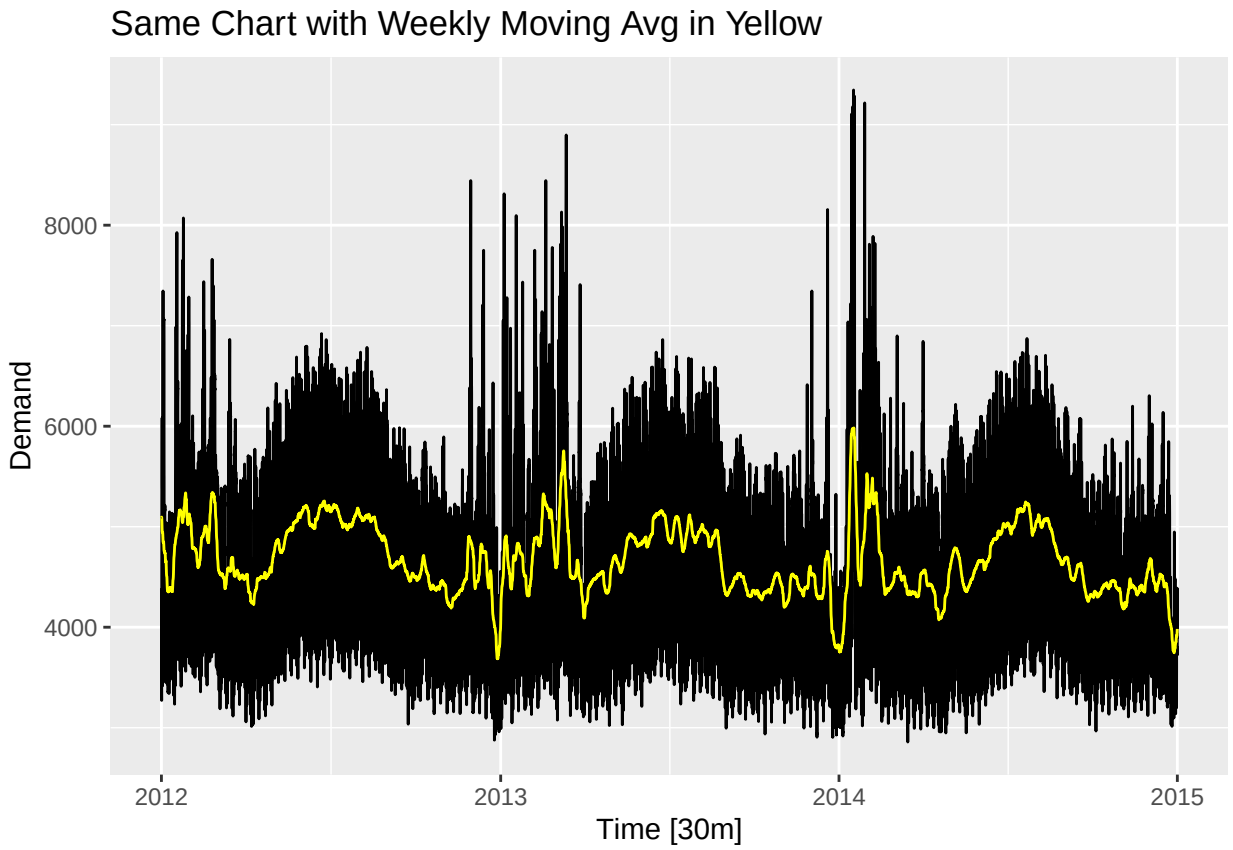
```
vic_elec %>%
  autoplot(Demand) +
  labs(title = "Electricity Demand in Victoria, Australia")
```



The winter months near the middle of each year show similar amounts of increase, presumably for heating demands. The daily spikes in air conditioning demand each summer look much harder to predict.

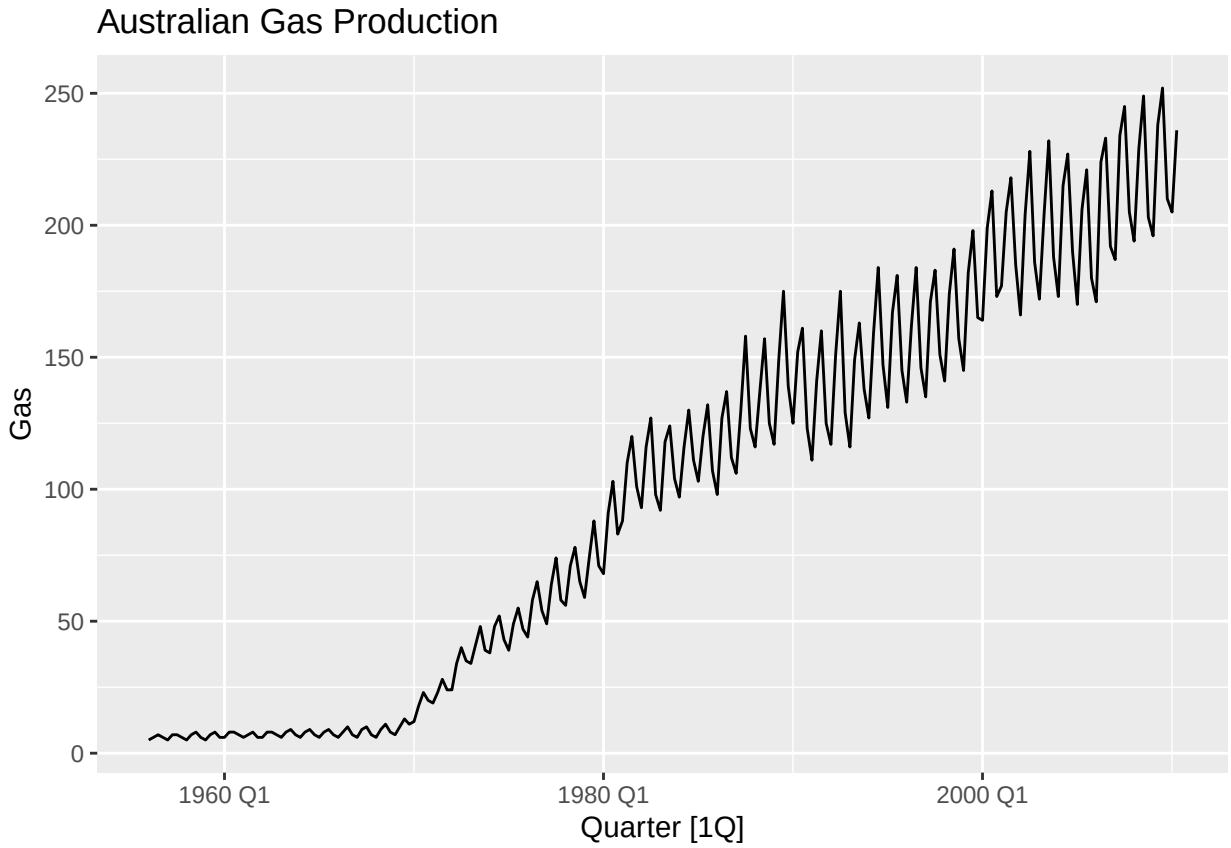
Before moving on to the next dataset, here is the electricity data from above with the weekly moving average overlaid to provide a better idea of how the demand is varying. Each yellow datapoint is averaged with the previous 168 and ensuing 168 half-hour measurements.

```
vic_elec %>%
  mutate(
    Weekly = slider::slide_dbl(Demand, mean,
                .before = 168, .after = 168, .complete = F)) %>%
  autoplot(Demand) +
  geom_line(aes(y = Weekly), colour = "yellow") +
  ggtitle("Same Chart with Weekly Moving Avg in Yellow")
```



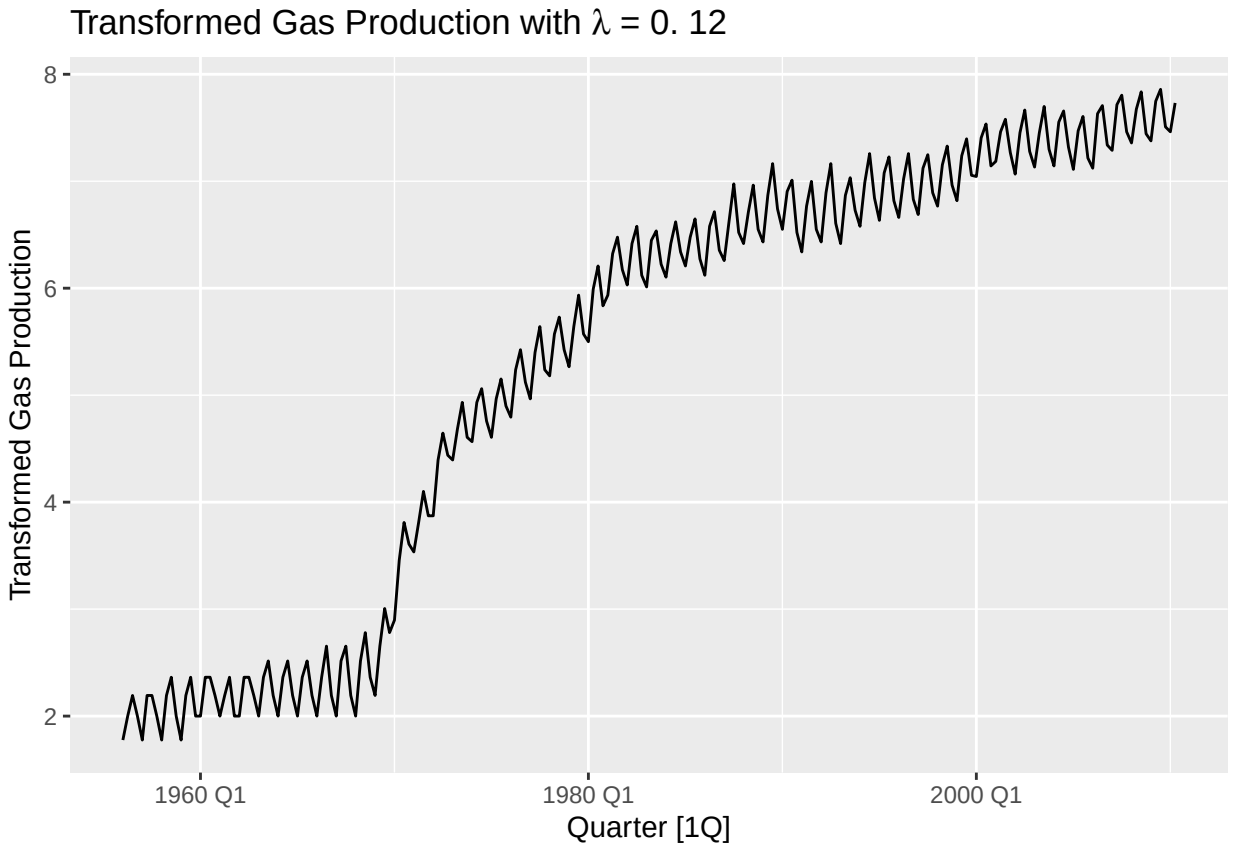
Gas production from `aus_production`

```
aus_production %>%
  autoplot(Gas) +
  labs(title = "Australian Gas Production")
```



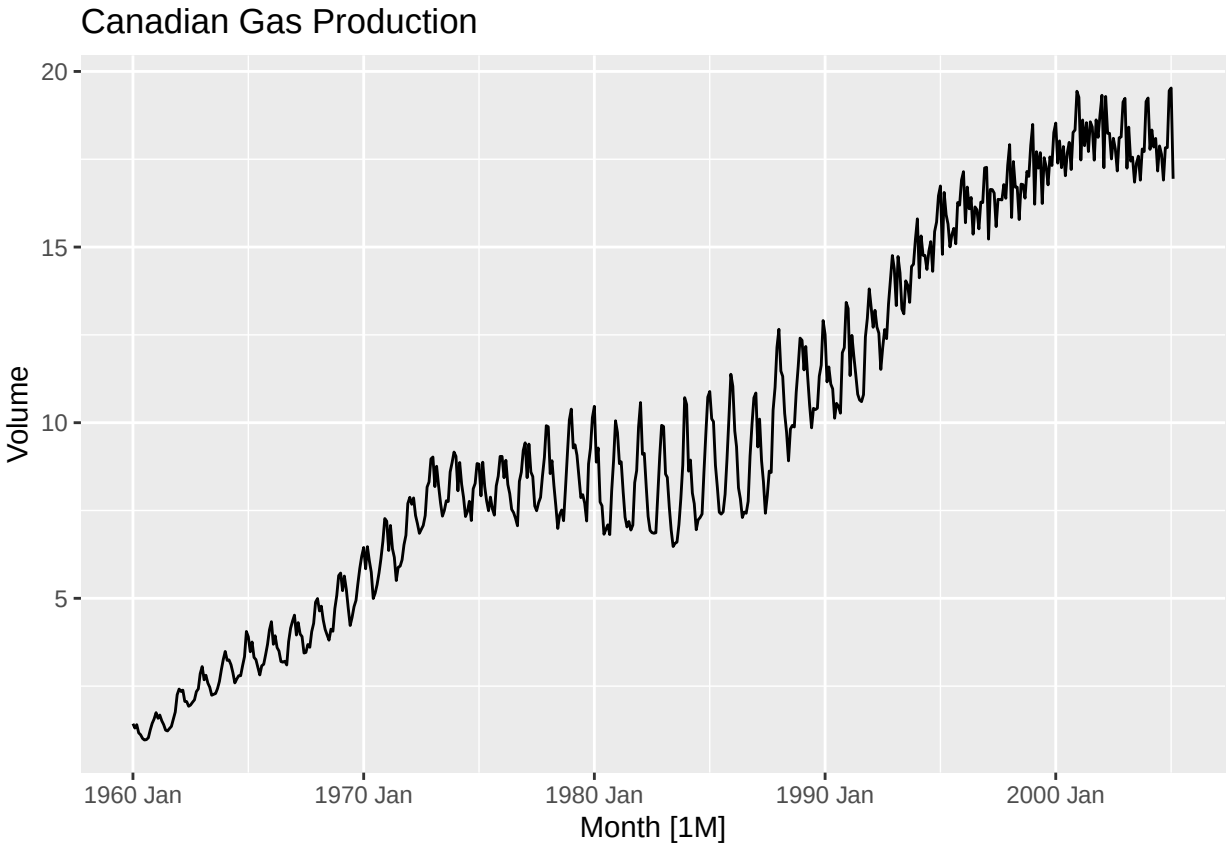
This is the example shown in part 3.1 of our book, where the Guerrero Method is used to find the Box-Cox transformation which evens out the magnitude of the seasonal variation, like this:

```
lambda <- aus_production %>%
  features(Gas, features = guerrero) %>%
  pull(lambda_guerrero)
aus_production %>%
  autoplot(box_cox(Gas, lambda)) +
  labs(y = "Transformed Gas Production",
       title = latex2exp::TeX(paste0(
         "Transformed Gas Production with $\\lambda$ = ",
         round(lambda,2))))
```



3) Why is a Box-Cox transformation unhelpful for the `canadian_gas` data?

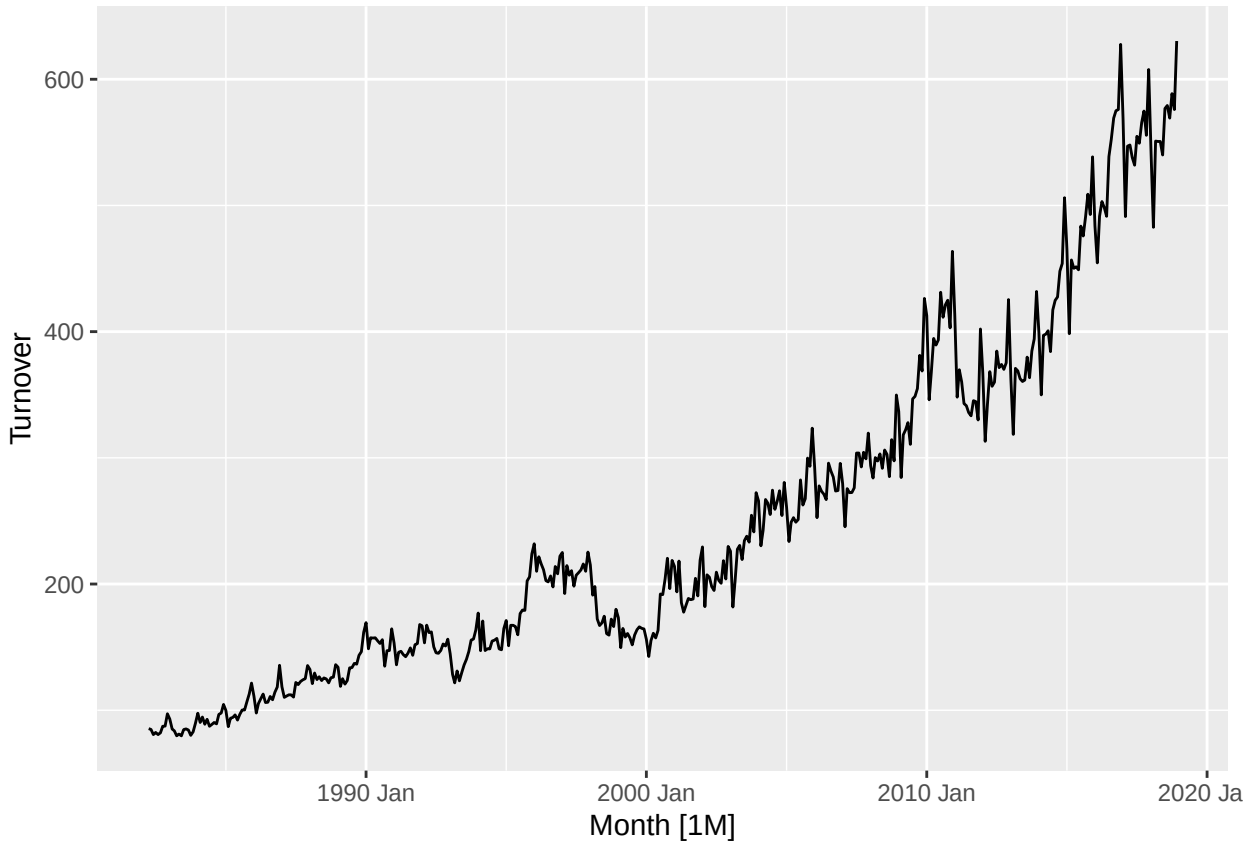
```
canadian_gas %>%
  autoplot(Volume) +
  labs(title = "Canadian Gas Production")
```



The magnitude of the seasonal variance changes independently from, or at least not monotonically with, how the Volume changes. It's interesting that in the 15 years (approximately 1973-1987) when the yearly production was most stable, the last 10 of those years have the least stable monthly production of any years in the chart.

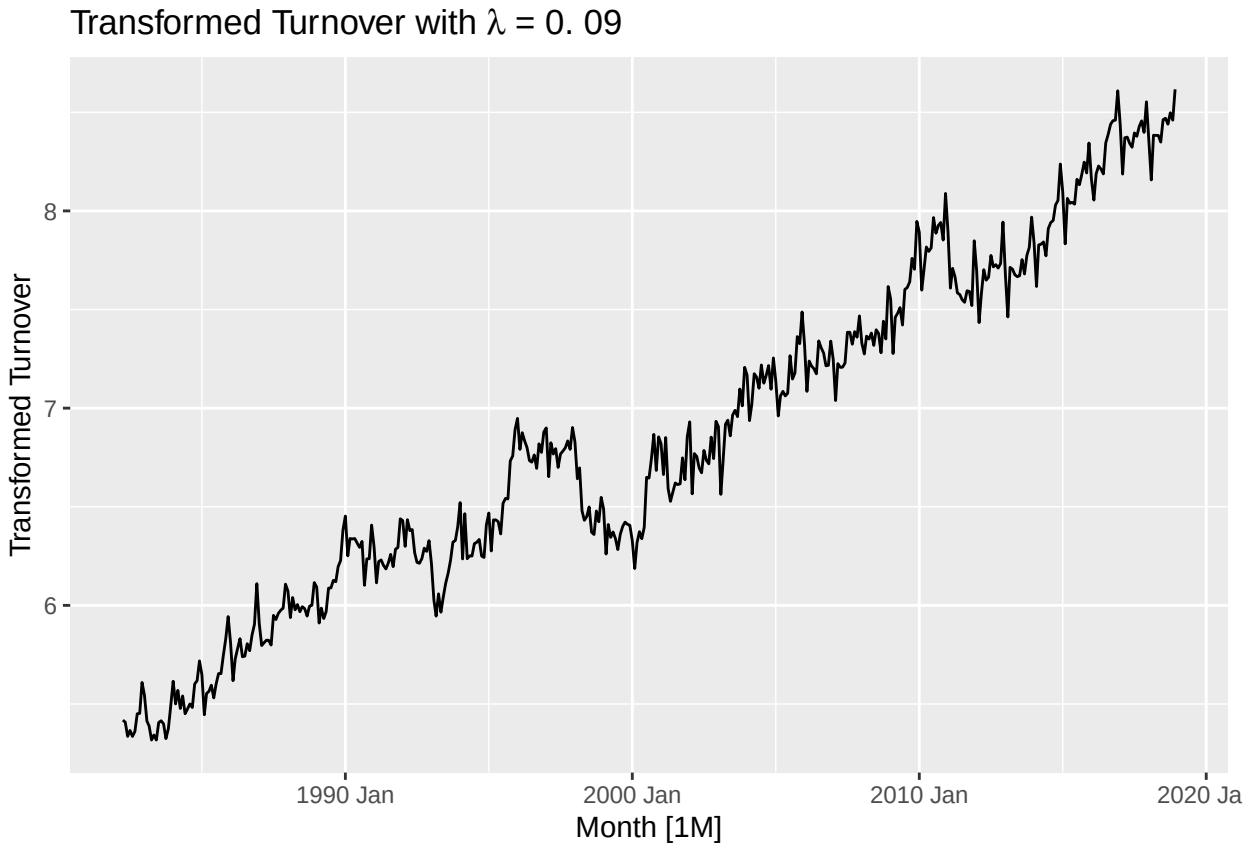
4) What Box-Cox transformation would you select for your retail data (from Exercise 8 in Section 2.10)?

```
set.seed(624)
myseries <- aus_retail %>%
  filter(`Series ID` == sample(aus_retail$`Series ID`,1))
myseries %>% autoplot(Turnover)
```



Again using the book's application of the Guerrero method, to even out the variance in the seasonal component of the data:

```
lambda <- myseries %>%
  features(Turnover, features = guerrero) %>%
  pull(lambda_guerrero)
myseries %>%
  autoplot(box_cox(Turnover, lambda)) +
  labs(y = "Transformed Turnover",
       title = latex2exp::TeX(paste0(
         "Transformed Turnover with $\\lambda$ = ",
         round(lambda,2))))
```



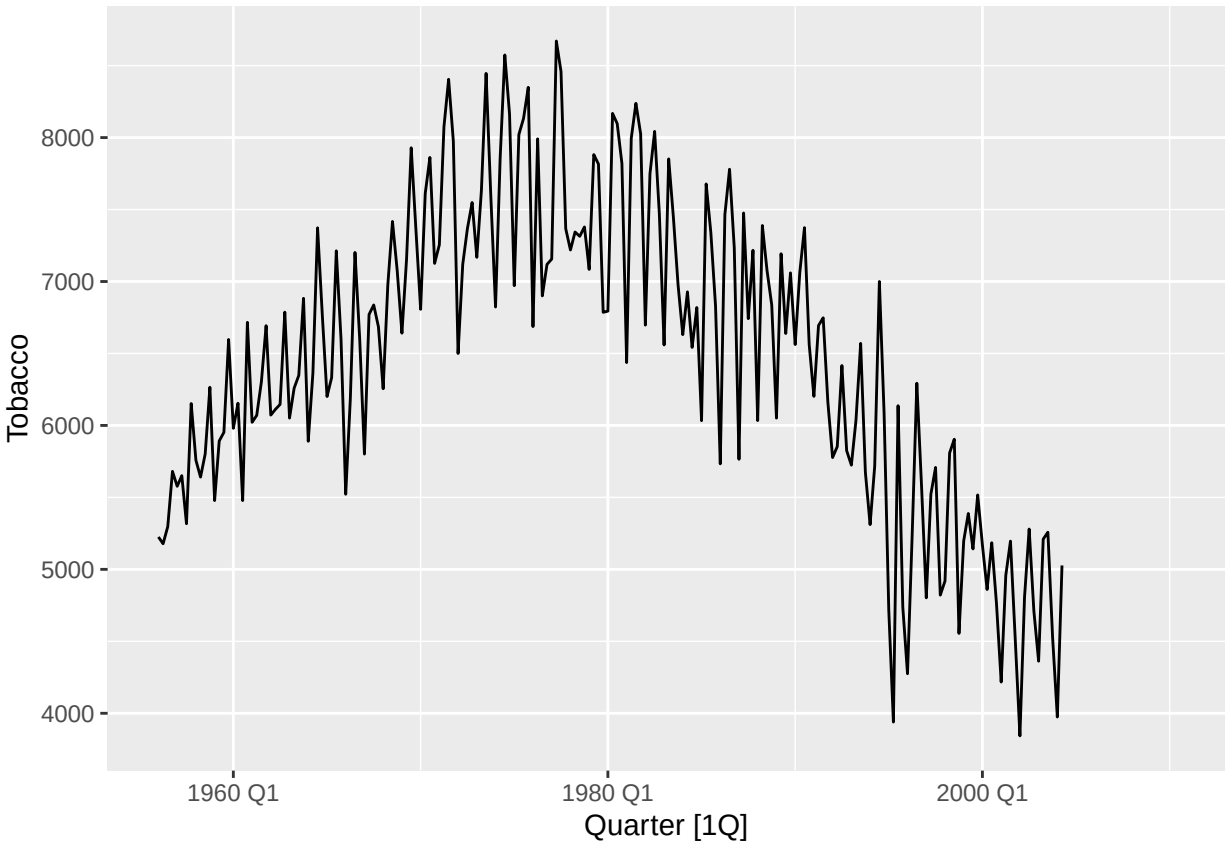
Another nice side effect of this Box-Cox power transformation is that it has made the overall trend more linear.

5) For the following series, find an appropriate Box-Cox transformation in order to stabilise the variance.

Tobacco from `aus_production`

```
aus_production %>%
  autoplot(Tobacco)
```

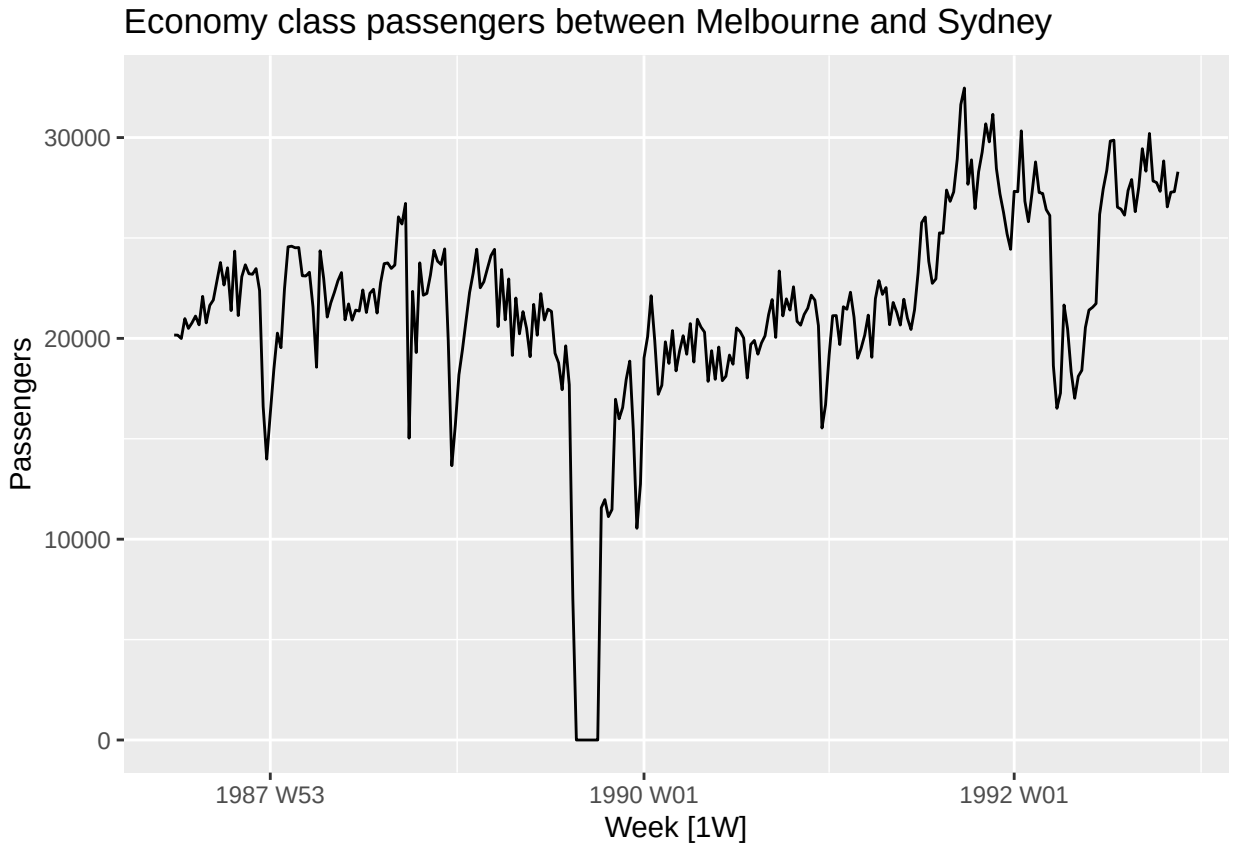
```
## Warning: Removed 24 row(s) containing missing values (geom_path).
```



I don't think I'd use a Box-Cox transform on the Tobacco data, since the seasonal variances don't seem tied to the value they vary around. 1995-98, for example, have similar production levels, but each successive year sees
decreasing seasonal variance.

Economy class passengers between Melbourne and Sydney from `ansett`

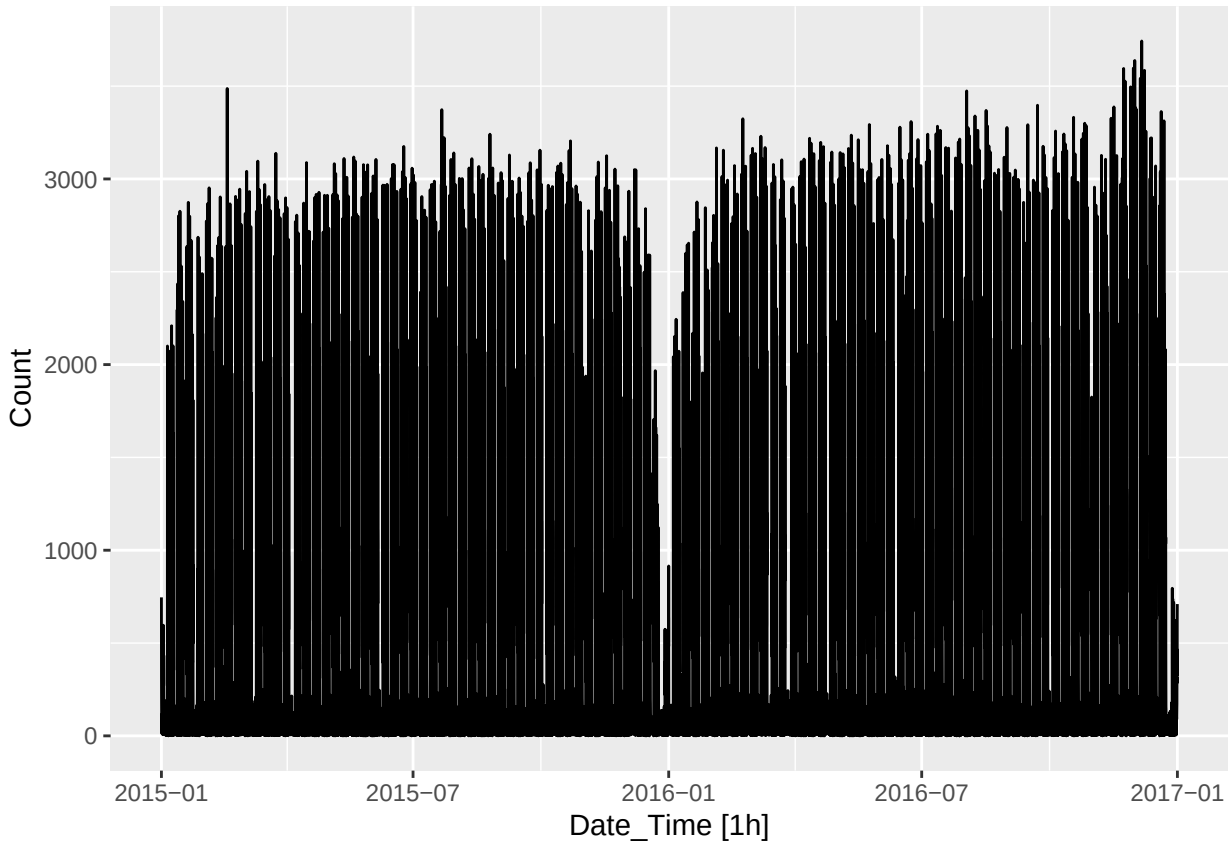
```
ansett %>%
  filter(Airports=="MEL-SYD") %>%
  filter(Class=="Economy") %>%
  autoplot(Passengers) +
  labs(title = "Economy class passengers between Melbourne and Sydney")
```



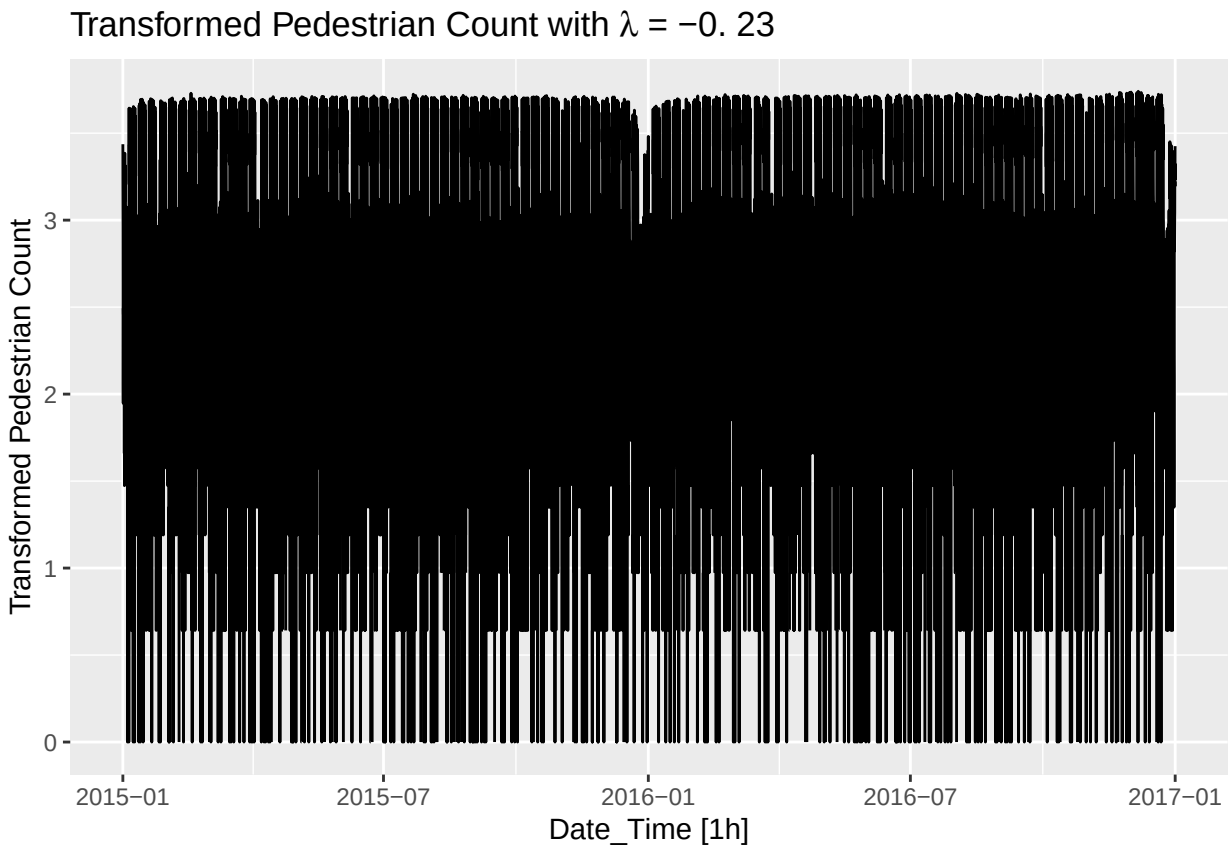
It's hard for me to see where a Box-Cox transformation would help with the biggest problem with the variance in those numbers: The large dropoffs that happen around New Years each year, as well as the patch of zeros in the second half of 1989.

Pedestrian counts at Southern Cross Station from `pedestrian`

```
pedestrian %>%
  filter(Sensor=="Southern Cross Station") %>%
  autoplot(Count)
```




```
lambda <- pedestrian %>%
  filter(Sensor=="Southern Cross Station") %>%
  features(Count, features = guerrero) %>%
  pull(lambda_guerrero)
pedestrian %>%
  filter(Sensor=="Southern Cross Station") %>%
  autoplot(box_cox(Count, lambda)) +
  labs(y = "Transformed Pedestrian Count",
       title = latex2exp::TeX(paste0(
         "Transformed Pedestrian Count with $\\lambda$ = ",
         round(lambda,2))))
```



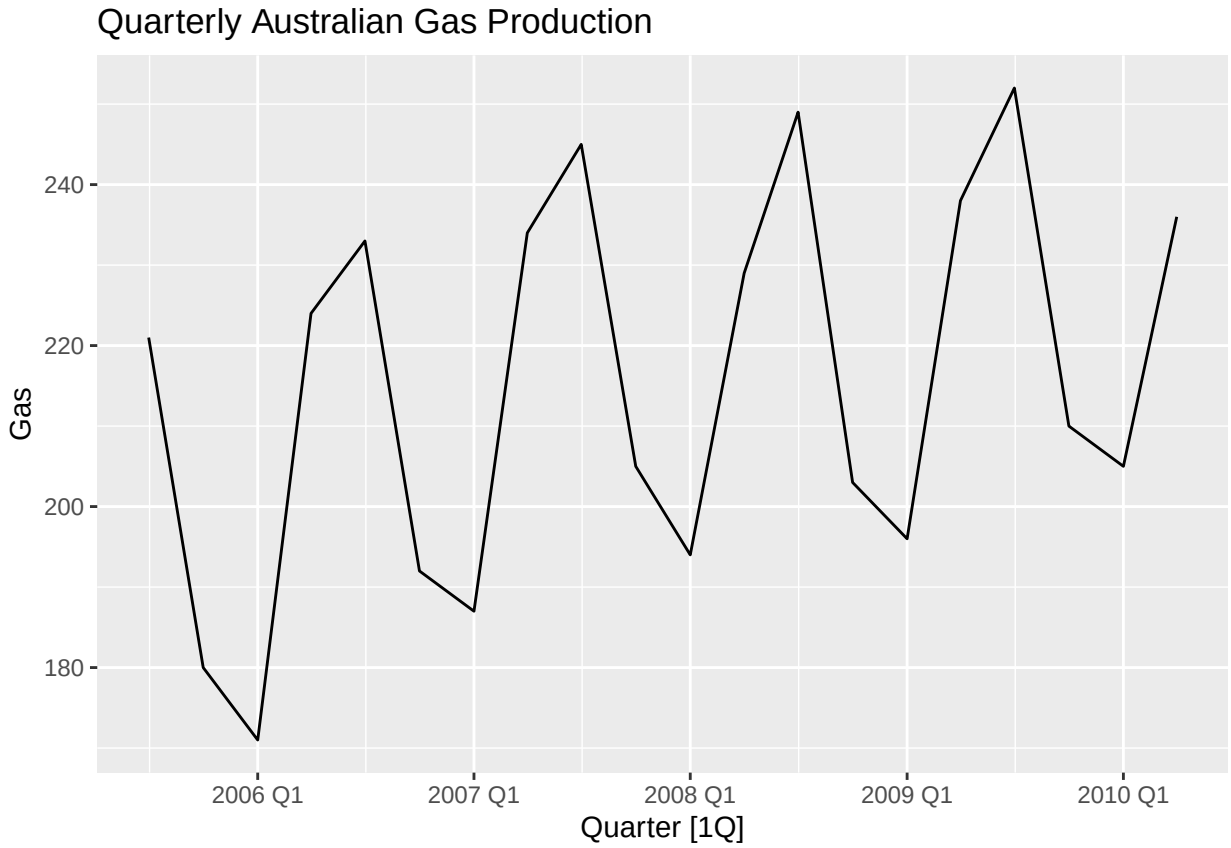
This Box-Cox transform does seem to have evened out the peaks and valleys of the daily cycle.

7) Consider the last five years of the Gas data from aus_production

```
gas <- tail(aus_production, 5*4) %>% select(Gas)
```

a) Plot the time series. Can you identify seasonal fluctuations and/or a trend-cycle?

```
gas %>%
  autoplot(Gas) +
  labs(title = "Quarterly Australian Gas Production")
```



The seasonal gas pattern shows high production in the winter and low production in the summer, suggesting the gas is hard to store. The trend for these five years is upward.

b) Use classical_decomposition with type=multiplicative to calculate the trend-cycle and seasonal indices

```
gas %>%
  model(classical_decomposition(Gas, type = "multiplicative")) %>%
  components() -> comps
```

Seasonal component:

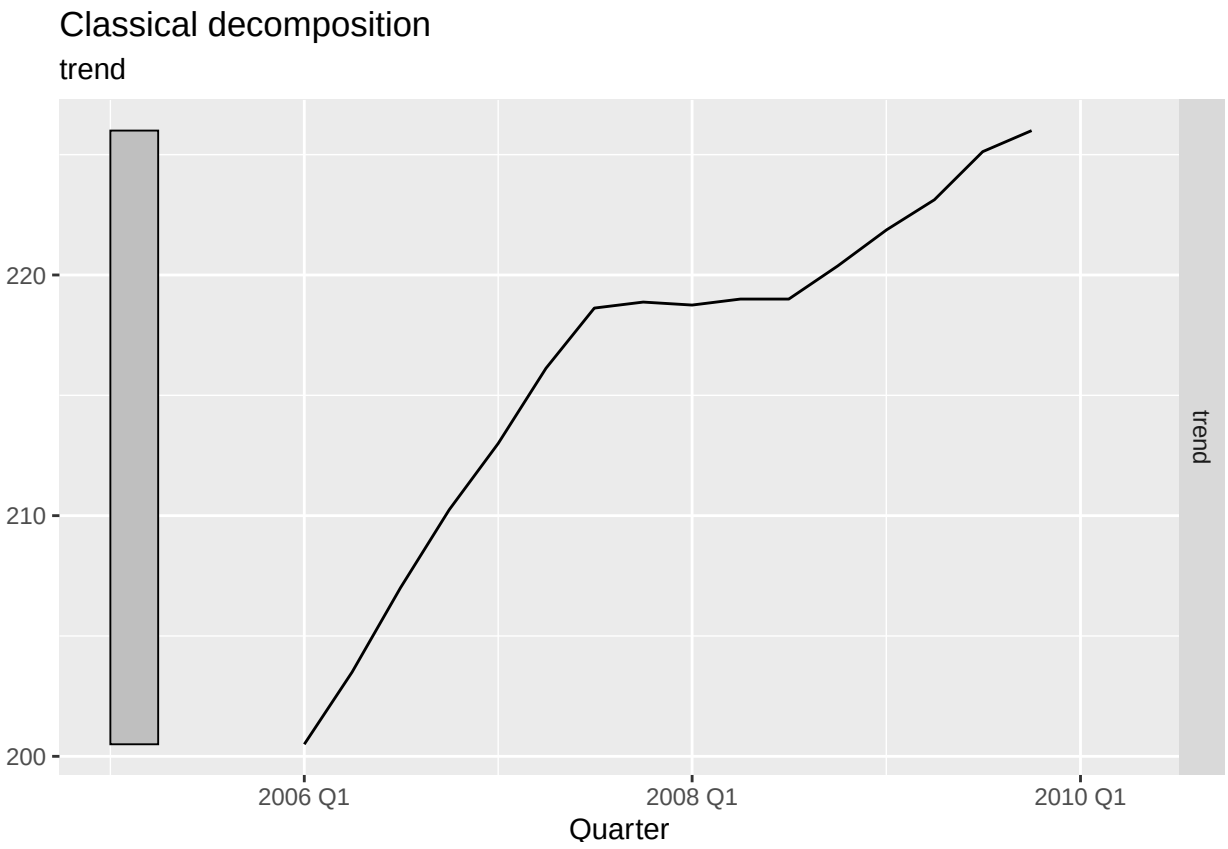
```
comps %>%
  select(seasonal) %>%
  head(5)
```

```
## # A tsibble: 5 x 2 [1Q]
##   seasonal Quarter
##      <dbl>   <qtr>
## 1    1.13  2005 Q3
## 2    0.925 2005 Q4
## 3    0.875 2006 Q1
## 4    1.07  2006 Q2
## 5    1.13  2006 Q3
```

Trend component:

```
autoplot(comps, trend)
```

```
## Warning: Removed 4 row(s) containing missing values (geom_path).
```



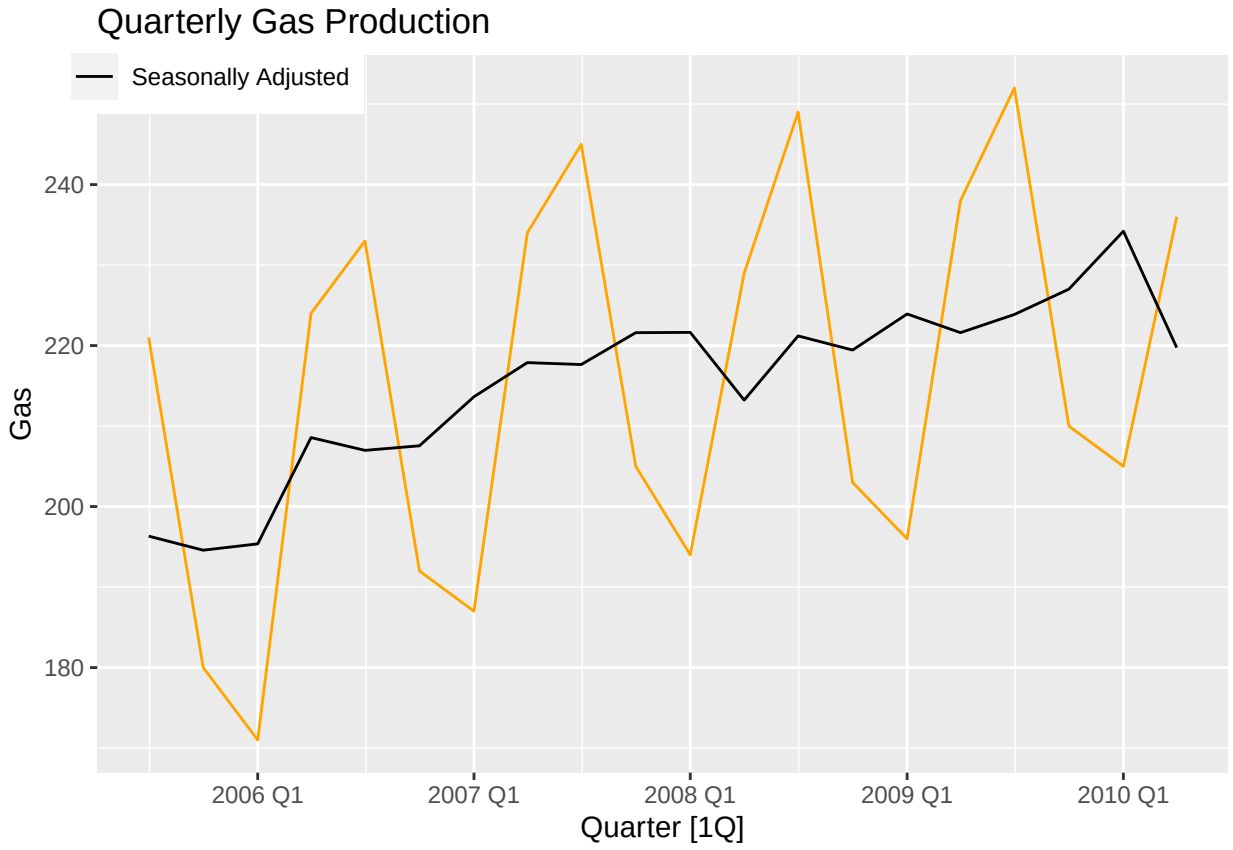
c) Do the results support the graphical interpretation from part a?

Q4 and Q1 have lower seasonal components whereas Q2 and Q3 have higher components, which supports the graphical interpretation.

The trend is indeed higher.

d) Compute and plot the seasonally adjusted data

```
# back out the seasonal factor
comps$seasAdj = comps$Gas / comps$seasonal
# remake a tsibble for easier plotting
comps %>%
  select(c(Gas, seasAdj)) %>%
  as_tsibble(index = Quarter) %>%
  autoplot(Gas, colour="orange") +
  geom_line(aes(y=seasAdj, colour = "Seasonally Adjusted")) +
  scale_color_manual(name='', values = c("Seasonally Adjusted" = "black")) +
  labs(title = "Quarterly Gas Production") +
  theme(legend.position = c(0.1, 1.0))
```



e) Change one observation to be an outlier (e.g., add 300 to one observation), and recompute the seasonally adjusted data. What is the effect of the outlier?

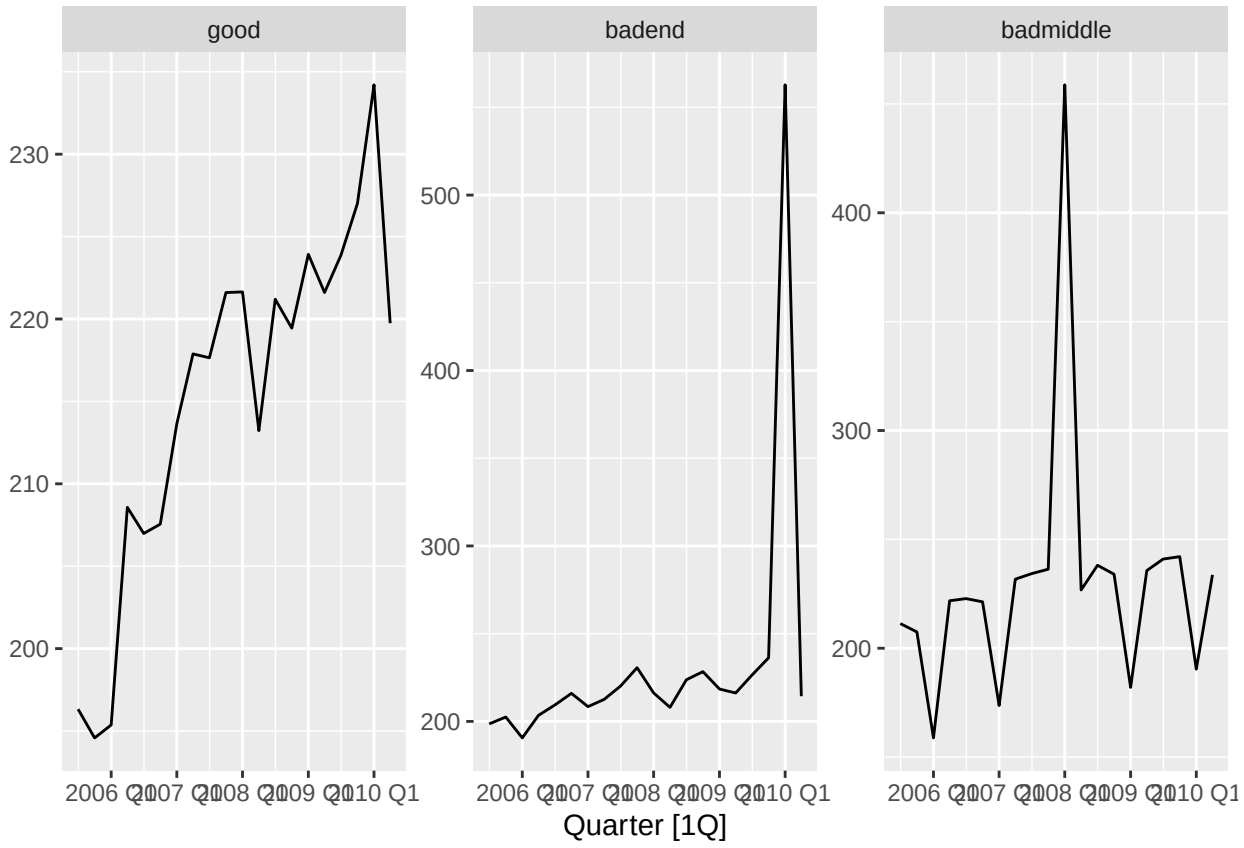
```
gas %>%   # add outliers at the same seasonal index, for better comparison
  mutate(BadMiddleGas = Gas + c(rep(0, 10), 300, rep(0, 9))) %>%
  mutate(BadEndGas = Gas + c(rep(0, 18), 300, 0)) -> badgas

badgas %>%
  model(classical_decomposition(BadMiddleGas, type = "multiplicative")) %>%
  components() -> badmidcomps

badgas %>%
  model(classical_decomposition(BadEndGas, type = "multiplicative")) %>%
  components() -> badendcomps

badgas$badmiddle = badmidcomps$season_adjust
badgas$badend = badendcomps$season_adjust
badgas$good = comps$season_adjust

autoplot(badgas, vars(good, badend, badmiddle))
```



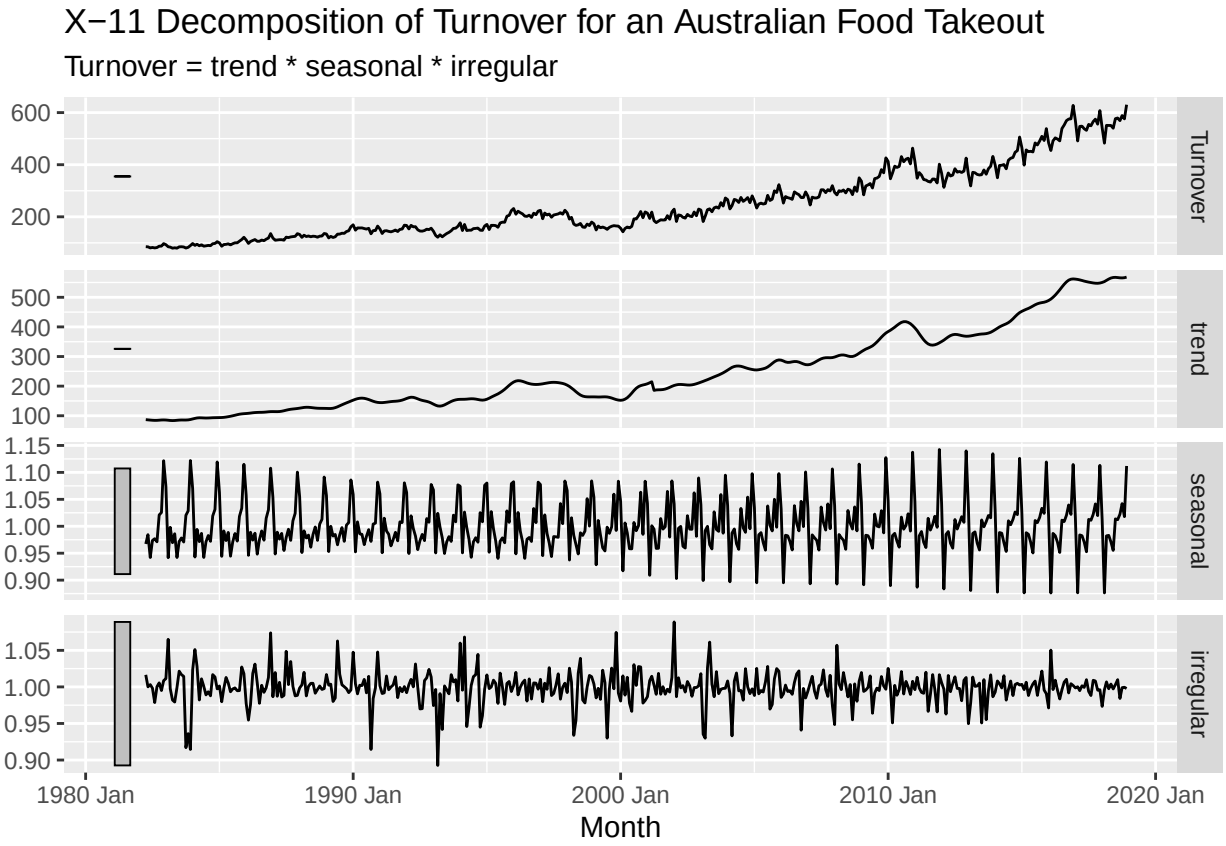
The outlier makes it very hard for the model to recognize seasonality, at least in a series that only has 5 years of cycles. That's evidenced by the fact that seasonality is still visible in the 2 corrupted data charts.

f) Does it make any difference if the outlier is near the end rather than in the middle of the time series?

The seasonality is less visible in the second chart, where the outlier is at the end, so I'd say it's better to have an outlier at the ends.

8) Recall your retail time series data (from Exercise 8 in Section 2.10). Decompose the series using X-11. Does it reveal any outliers, or unusual features that you had not noticed previously?

```
# library(seasonal)
myseries %>%
  model(x11 = X_13ARIMA_SEATS(Turnover ~ x11())) %>%
  components() -> x11_dcmp
autoplot(x11_dcmp) +
  labs(title =
    "X-11 Decomposition of Turnover for an Australian Food Takeout")
```



The seasonality grew more extreme as time advanced, which is why $\lambda = .09$ worked in the Box-Cox transform earlier. But this X-11 seasonal decomposition also shows that
things were actually the opposite in the 1980's, when the seasonal element decreased
as Turnover increased.

The bottom panel showing the "remainder" of the decomposition shows how small these remainders were for the last ten years of data, compared to earlier.

9) Figures 3.19 and 3.20 show the result of decomposing the number of persons in the civilian labour force in Australia each month from February 1978 to August 1995.

Write about 3–5 sentences describing the results of the decomposition. Pay particular attention to the scales of the graphs in making your interpretation.

Much like with the takeout food Turnover shown above, these labour numbers are
primarily moving with an upward trend, although there is a clear seasonal component.
The August low points and December high points have become more extreme each year,
which suggests that a logarithmic or Box-Cox transform might even those cycles out.

For some reason, March was a month of especially high employment in the 1980's
and then became much less so in the 1990's.

Is the recession of 1991/1992 visible in the estimated components?

The trend and season components seem to have avoided the head-fake by the recession,
and consequently the model uses the remainder component to explain it. I don't know the exact dates of the recession, but only 1991's labour numbers appear to have
been affected, not 1992's.