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

Kevin Kirby

2024-09-12

The below are answers to exercises 3.1, 3.2, 3.3, 3.4, 3.5, 3.7, 3.8 and 3.9 from section 3.7 of the [Forecasting: Principles and Practice (3rd ed)] (https://otexts.com/fpp3/graphics-exercises.html). This is homework one of the DATA 624 class "Predictive Analytics." Unless otherwise noted, all datasets used below are from the fpp3 package that's owned and maintained by the book's authors.

First, I'll load the required libraries:

```r
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()
```

```r
library(ggplot2)
library(scales)
library(plotly)
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
```

```
##
##     layout
```

```
library(patchwork)
library(dplyr)
library(seasonal)
```

```
##
## Attaching package: 'seasonal'

## The following object is masked from 'package:tibble':
##
##     view
```

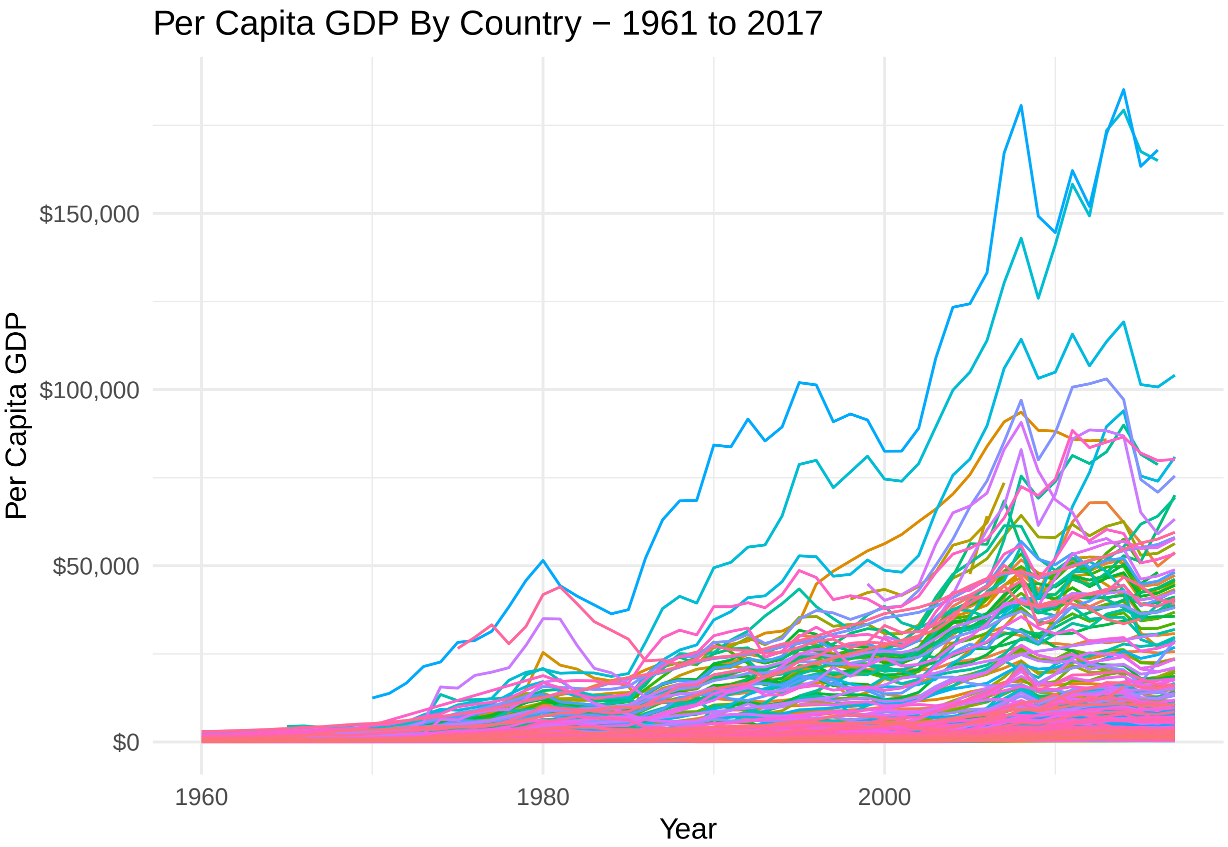
Exercises

3.1 Plot Per Capita GDP By Country from global_economy

The first exercise asks for a plot of Per Capita GDP by Country. To do this, I will need to divide GDP by Population and remove missing GDP values.

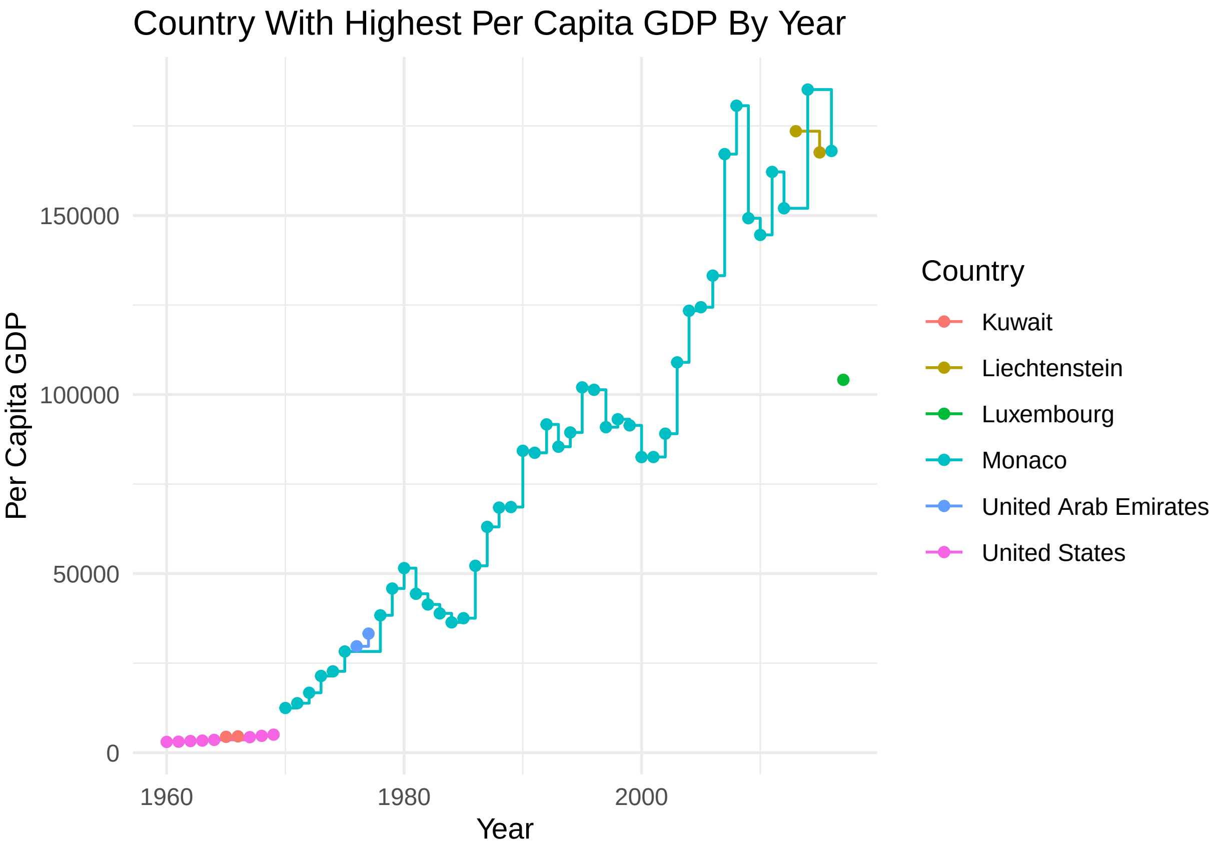
First, I'll create a new dataset that has per capita GDP with country and year. The rest isn't needed for this plot makes the overall dataset smaller without removing required information.

```
per_cap_gdp <- global_economy %>%
  mutate(per_cap_gdp = GDP / Population) %>%
  select(Country, Year, per_cap_gdp) %>%
  filter(!is.na(per_cap_gdp))

ggplot(per_cap_gdp, aes(x = Year, y = per_cap_gdp, color = Country)) +
  geom_line() +
  labs(title = "Per Capita GDP By Country - 1961 to 2017",
       x = "Year",
       y = "Per Capita GDP") +
  scale_y_continuous(labels = scales::dollar_format(scale = 1, prefix = "$", big.mark = ",")) +
  theme_minimal() +
  theme(legend.position = "none")
```



To determine the country with the highest GDP per capita and assess trends overtime, I used a step plot chart that colors the country and shows the GDP for that year.

```r
highest_gdp <- per_cap_gdp %>%
  index_by(Year) %>%
  slice_max(order_by = per_cap_gdp, n = 1, with_ties = FALSE) %>%
  arrange(Year) %>%
  select(Year, Country, per_cap_gdp) %>%
  mutate(prev_country = lag(Country)) %>%
  filter(is.na(prev_country) | Country != prev_country) %>%
  select(-prev_country)

ggplot(highest_gdp, aes(x = Year, y = per_cap_gdp, color = Country)) +
  geom_step() +
  geom_point() +
  labs(title = "Country With Highest Per Capita GDP By Year",
       x = "Year",
       y = "Per Capita GDP",
       color = "Country") +
  theme_minimal()
```



Country with the highest GDP for 2017 (the most recent year in the dataset): Luxemborg, the first year it's held the title. It's a country of under 100,000 people with per capita GDP of over $100,000.

Monaco had the longest run on top, stretching from 1978 to 2012. The United States hasn't held the title since 1969, which doesn't surprise me. American power comes from the sheer amount of money in total we have and, relative to our country size, our overall not bad standard of living.

3.2 - Graphs, Dang Graphs, and Statistics

The next exercise says: "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.
- Slaughter of Victorian "Bulls, bullocks and steers" in aus_livestock.
- Victorian Electricity Demand from vic_elec.
- Gas production from aus_production."

US GDP from global_economy

I created two charts side by side: one without log transformation and one with it. GDP tends to grow exponentially and a log transformation is appropriate when looking to create a more rational curve.

"'{r-usgdp}

us_gdp <- global_economy %>% filter(Country == "United States")

ggplot(us_gdp, aes(x = Year, y = GDP)) + geom_line() + scale_y_continuous(labels = comma) + labs(title = "United States Yearly GDP - 1961-2017", y = "Dollars")

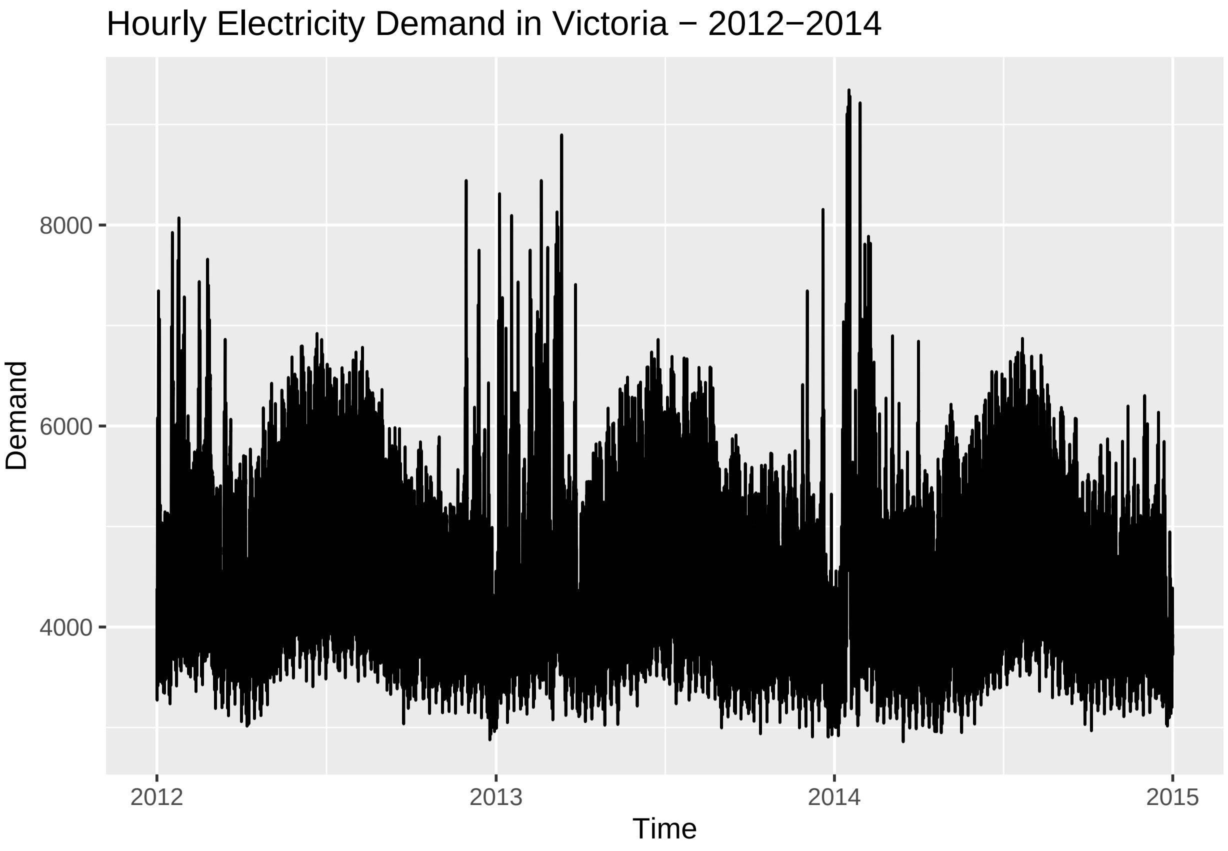
ggplot(us_gdp, aes(x = Year, y = GDP)) + geom_line() + scale_y_log10() + labs(title = "Log Scaled United States Yearly GDP - 1961-2017", y = "Dollars")

~~~
### Slaughter of Victorian "Bulls, bullocks and steers" in aus_livestock

There's some seasonality at play here but decided against any transformation. This type of change is ei
* Ecological changes, where there are less overall to be killed
* Policy driven: Australia implemented a series of policies that forced the change, either directly or i


```{r-victorian-slaughter}

aus_livestock %>%
  filter(Animal == "Bulls, bullocks and steers", State == "Victoria") %>%
  ggplot(aes(x = Month, y = Count)) +
  geom_line() +
  labs(title = "Slaughter of Victorian Bulls, Bullocks and Steers - July 1976 to December 2018 ", y = "I
~~~

Victorian Electricity Demand from vic_elec

Those spikes in demand at the start of each year line up with the peak of Australian summer in January. Otherwise, this data is stable and doesn't really require a transformation to perform objective analysis.

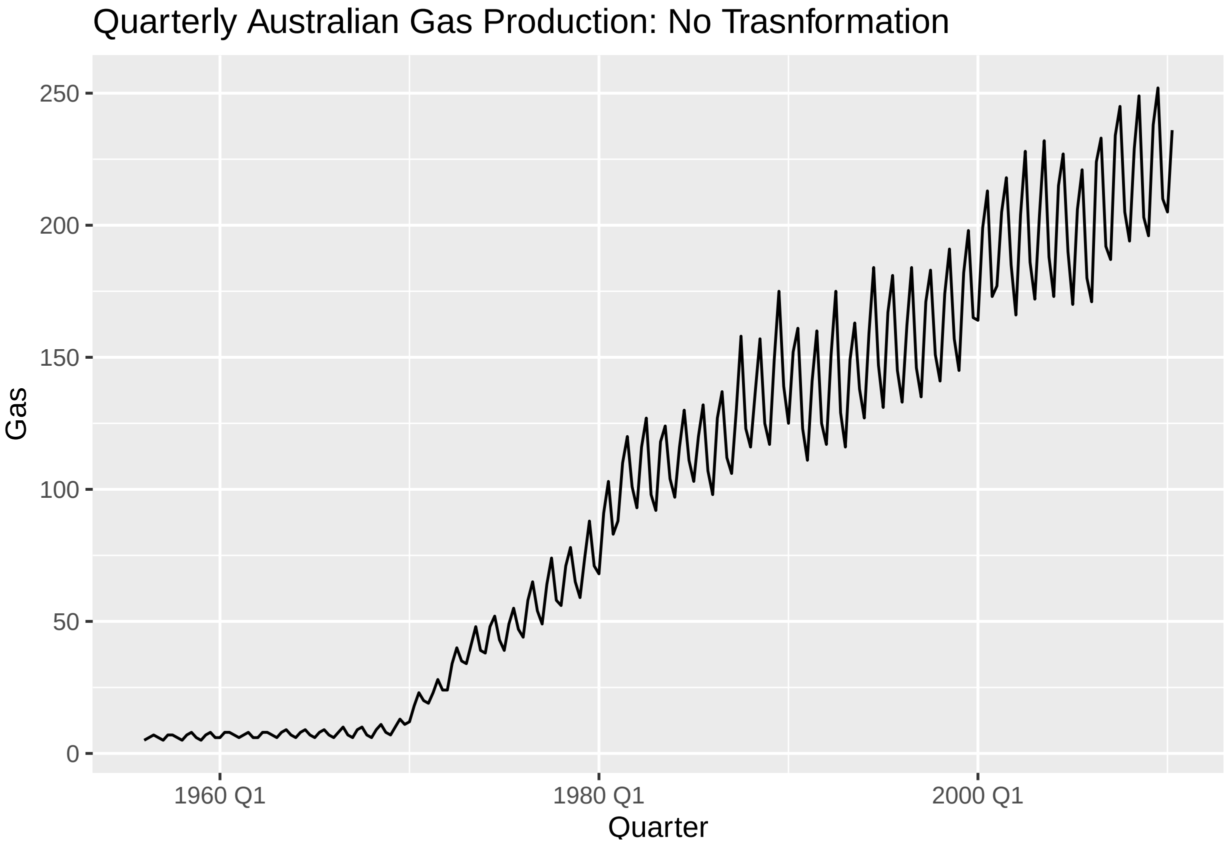
```r
ggplot(vic_elec, aes(x = Time, y = Demand)) +
  geom_line() +
  labs(title = "Hourly Electricity Demand in Victoria - 2012-2014", y = "Demand")
```



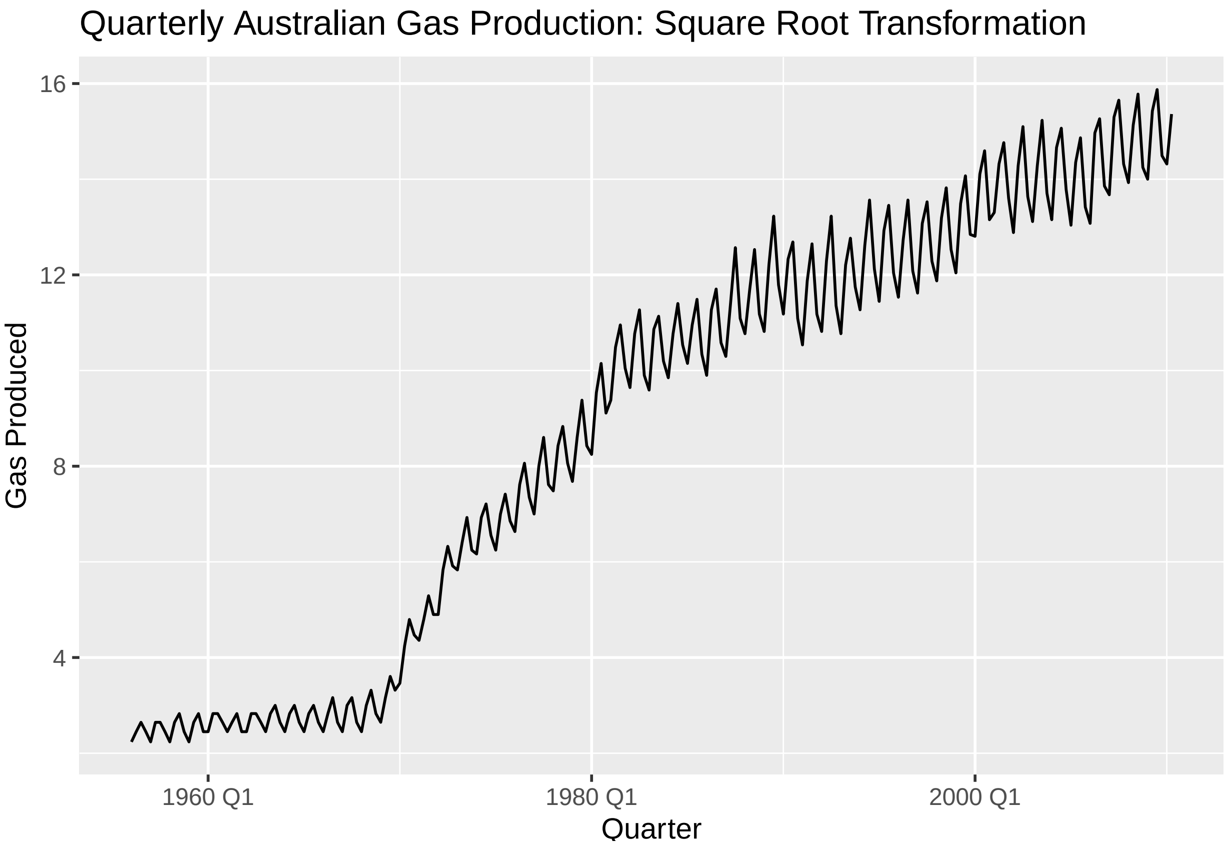
Gas production from aus_production

The next two charts show 1956 to 2010 quarterly Australian gas production. The first is just a regular plot with no transformation. The second is the same dataset with a square root transformation. A log transformation can be a pretty harsh and aggressive change on a dataset so I went with square root to stablize variance. You can see how what appears to be a solidly upward trend has actually been starting to level off.

```r
ggplot(aus_production, aes(x = Quarter, y = Gas)) +
  geom_line() +
  labs(title = "Quarterly Australian Gas Production: No Trasnformation", y = "Gas")
```



```r
ggplot(aus_production, aes(x = Quarter, y = sqrt(Gas))) +
  geom_line() +
  labs(title = "Quarterly Australian Gas Production: Square Root Transformation", y = "Gas Produced")
```

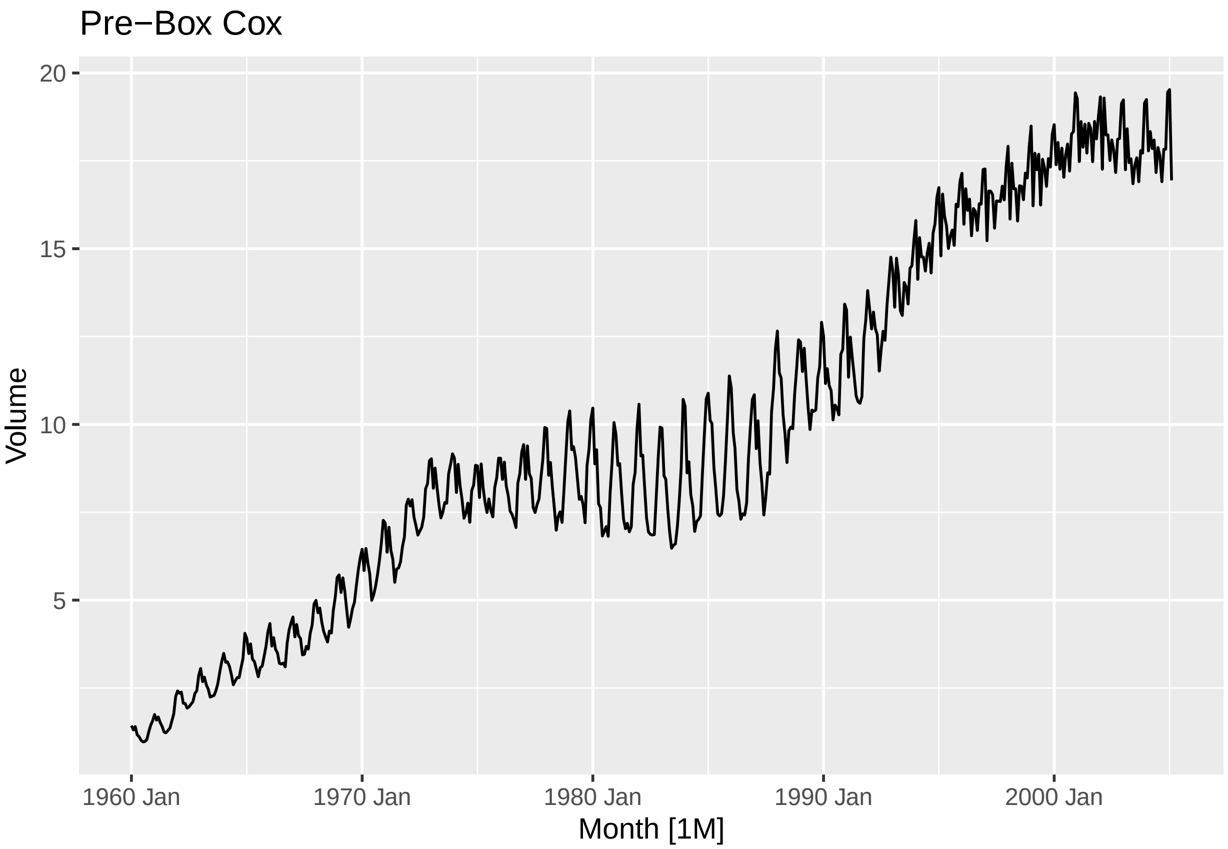


3.3 - Candian Gas Box Plot Analysis

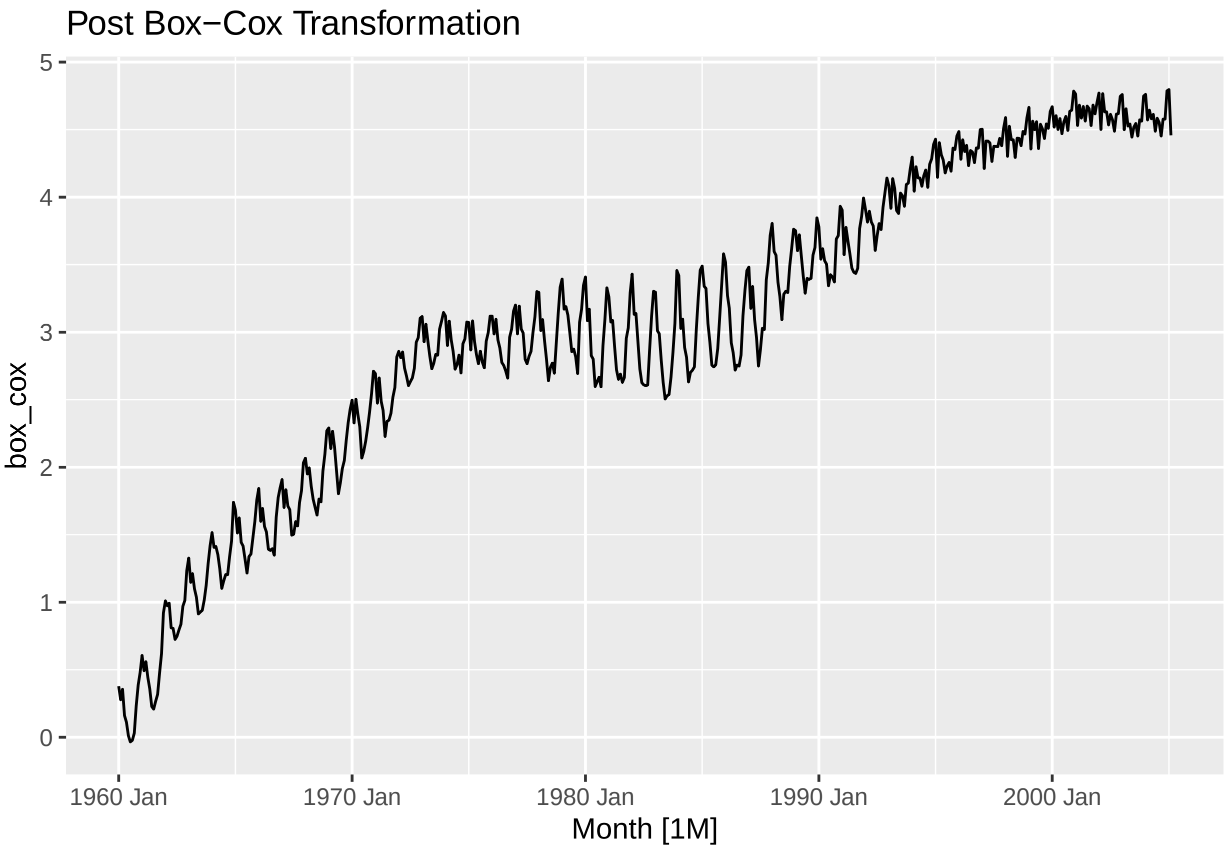
This exercise starts with the premise that Box-Cox-Transformation is "unhelpful for the canadian_gas data" and asks me why. To validate or reject the premise of the question (how do I personally know they're unhelpful?), I will plot both and then assess.

Both charts plot January 1960 to February 2005 monthly gas production in Canada.

```
cg_boxcox <- canadian_gas %>%
  as_tsibble() %>%
  mutate(box_cox = box_cox(Volume, lambda = 0.3))

canadian_gas %>%
  autoplot(Volume) +
  ggtitle("Pre-Box Cox")
```



```
cg_boxcox %>%
  autoplot(box_cox) +
  ggtitle("Post Box-Cox Transformation")
```



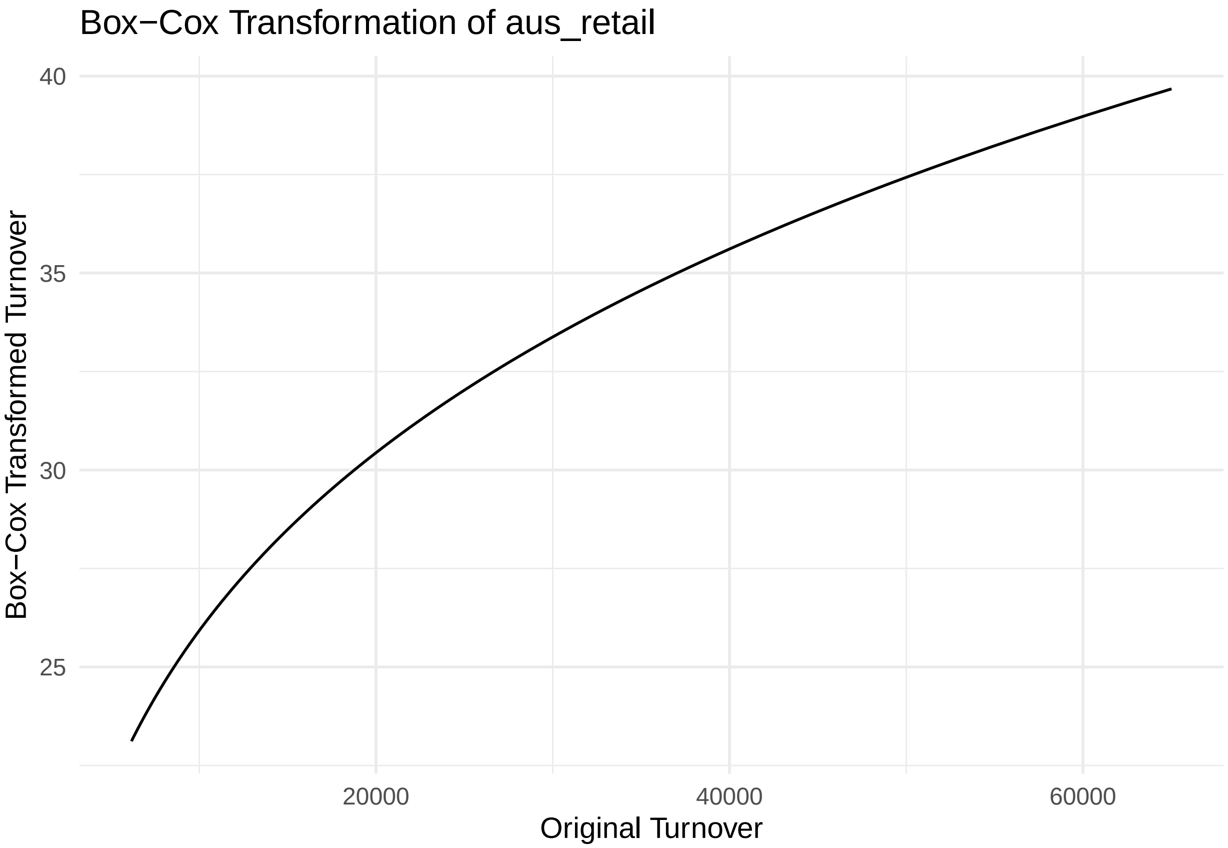
The transformation chart is not useful because the baseline data is predictable. Gas volume production has gone up over time at a steady clip, which lines up with Canada discovering more and more gas fields to dig up.Additionally, there isn't chaotic variance and there is seasonality. A separate investigation should be done into why the seaonality patterns started to changein the late 90s.

3.4 - Picking Box-Cox for aus_retail

This exercise asks me what Box-Cox transformation I would pick for the aus_retail dataset. I would pick the Guerro Method because it automatically picks an optimal Lambda for you. This is in the same spirit as the Adams Optimizer for gradient descent I use in deep learning, allowing me to have the model automatically adapt the learning rate.

Here's how the overall lambda for aus_retail can be generated. The value ends up being 0.196

```
aus_retail_la <- aus_retail %>%
  summarise(Turnover = sum(Turnover)) %>%
  features(Turnover, features = guerrero) %>%
  pull(lambda_guerrero)

aus_retail_transformed <- aus_retail %>%
  summarise(Turnover = sum(Turnover)) %>%
  mutate(Turnover_BoxCox = box_cox(Turnover, lambda = aus_retail_la))


aus_retail_transformed %>%
  ggplot(aes(x = Turnover, y = Turnover_BoxCox)) +
  geom_line() +
  labs(
    title = "Box-Cox Transformation of aus_retail",
    x = "Original Turnover",
    y = "Box-Cox Transformed Turnover"
  ) +
  theme_minimal()
```



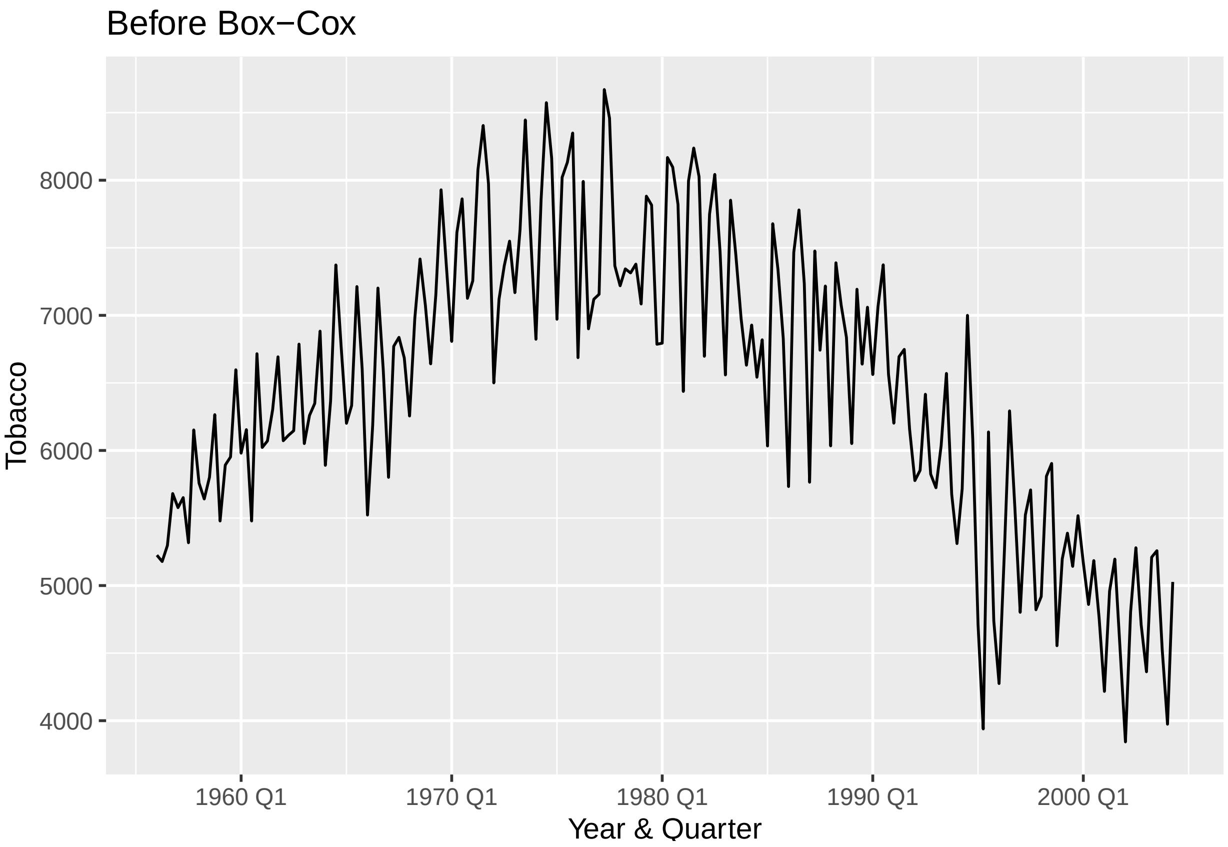
3.5 - Variance Stablization Through Box-Cox Transformation

This exercise asks me to find an appropriate Box-Cox transformation to stabilize the variances for:

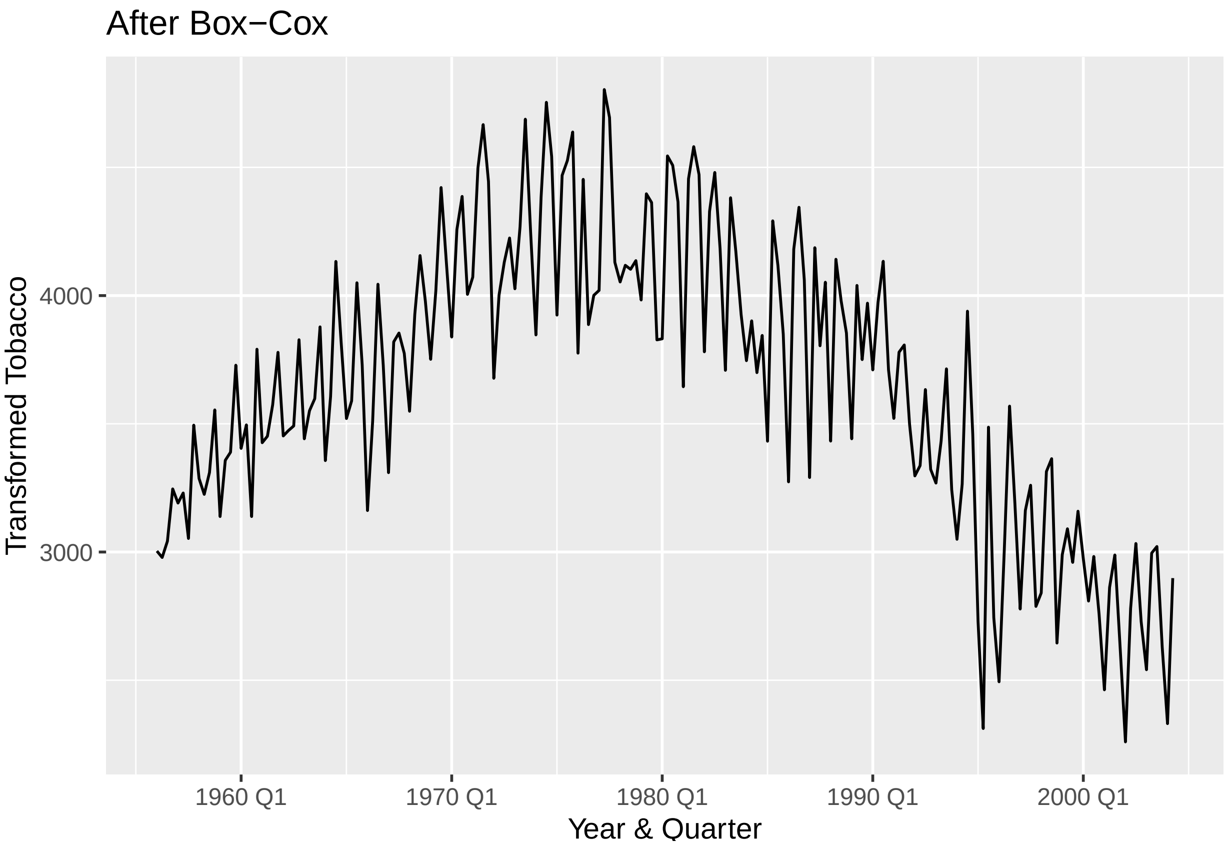
- Tobacco from aus_production
- Economy class passengers between Melbourne and Sydney from ansett
- Pedestrian counts at Southern Cross Station from pedestrian

Tobacco from aus_production

```
autoplot(aus_production %>% filter(!is.na(Tobacco)), Tobacco) +
  labs(title = "Before Box-Cox",
       x = "Year & Quarter",
       y = "Tobacco")
```

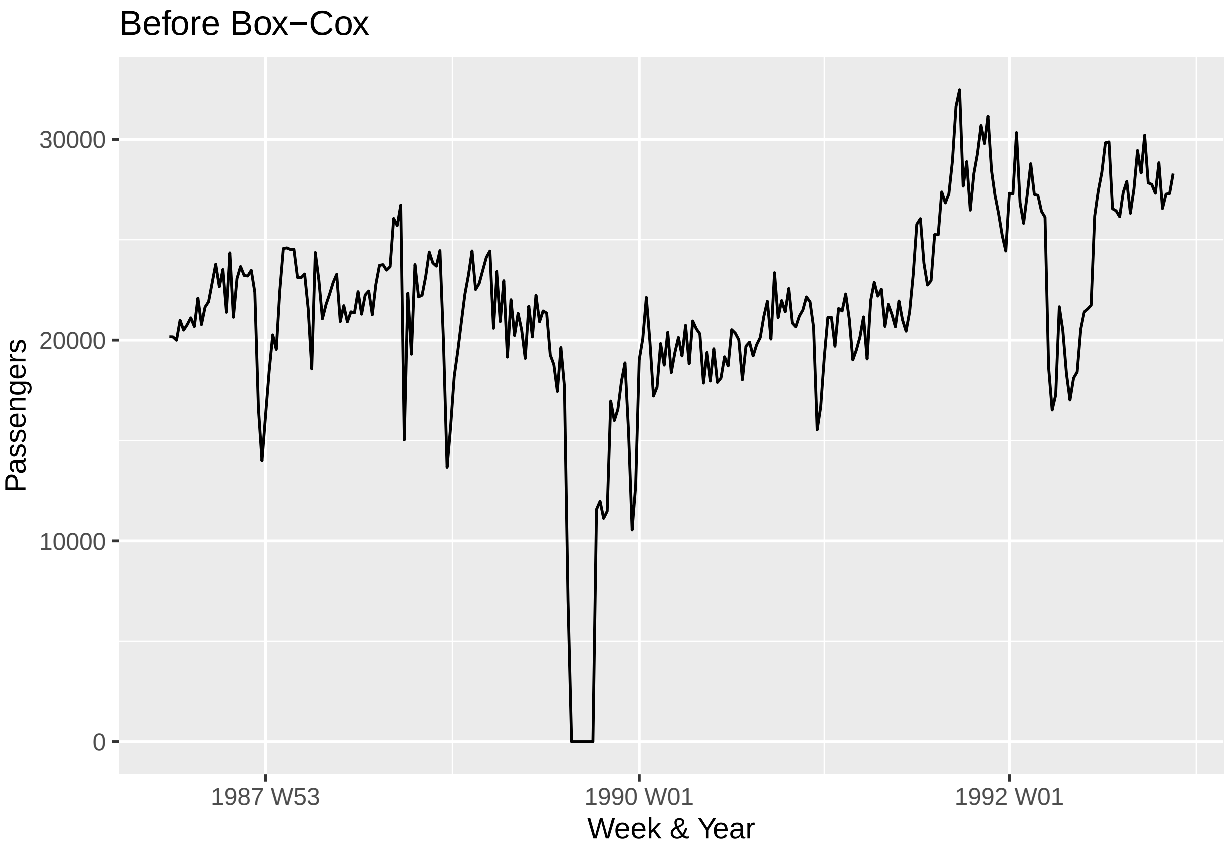


```
aus_tob_la <- aus_production %>%
  filter(!is.na(Tobacco)) %>%
  summarise(Tobacco = sum(Tobacco)) %>%
  features(Tobacco, features = guerrero) %>%
  pull(lambda_guerrero)

aus_tob_boxcox <- aus_production %>%
  filter(!is.na(Tobacco)) %>%
  as_tsibble() %>%
  mutate(box_cox = box_cox(Tobacco, lambda = aus_tob_la))

autoplot(aus_tob_boxcox, box_cox) +
  labs(title = "After Box-Cox",
       x = "Year & Quarter",
       y = "Transformed Tobacco")
```

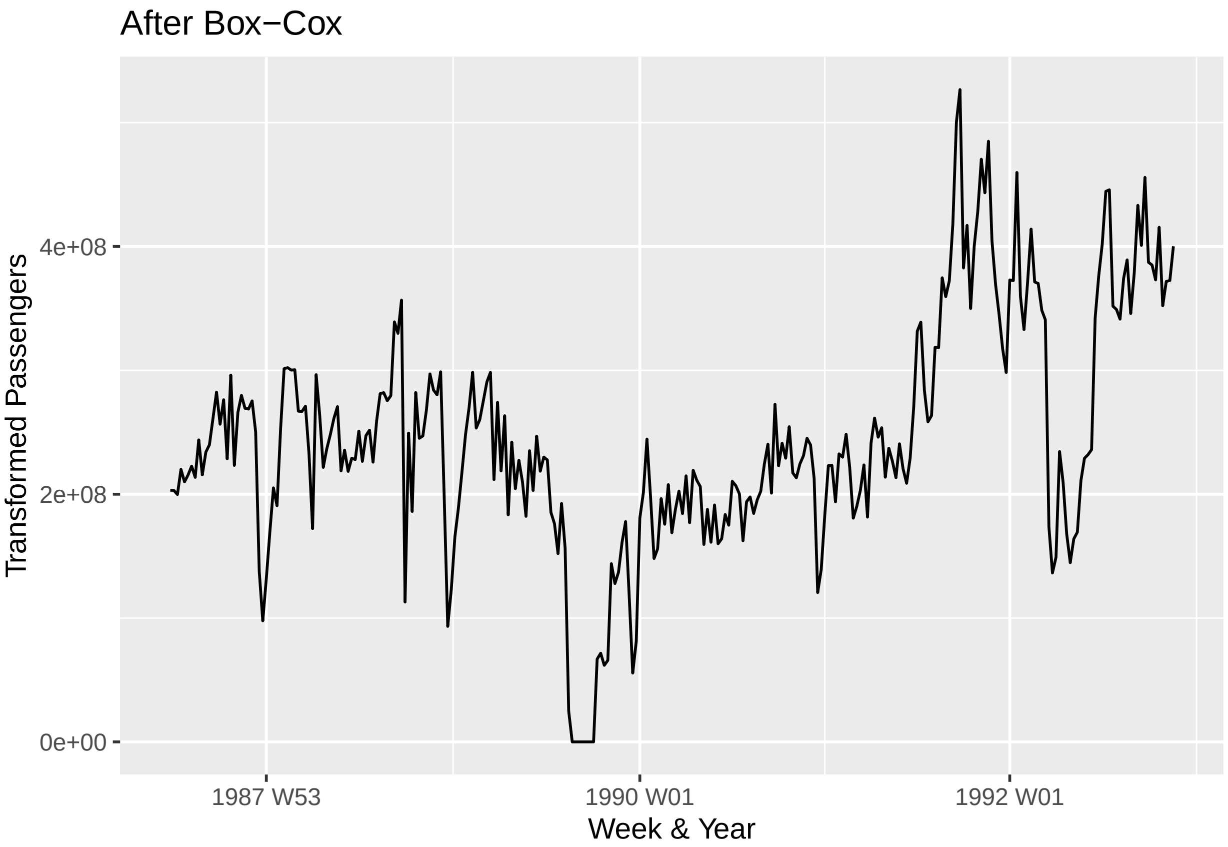


Economy class passengers between Melbourne and Sydney from ansett

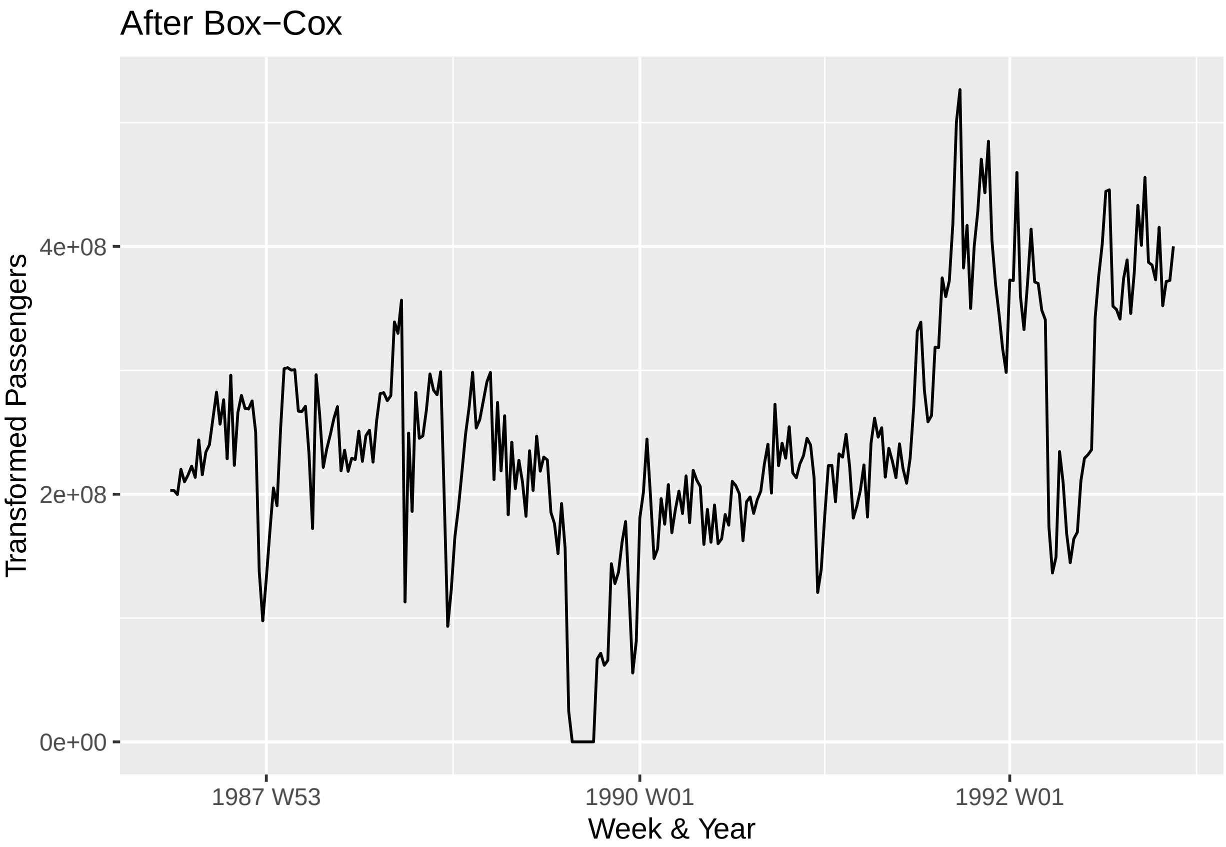
```
ansett_na <- ansett %>%
  drop_na(Passengers)

autoplot(ansett_na %>%
           filter(Airports == "MEL-SYD", Class == "Economy"), Passengers) +
  labs(title = "Before Box-Cox",
       x = "Week & Year",
       y = "Passengers")
```



```
ansett_la <- ansett_na %>%
  filter(Airports == "MEL-SYD", Class == "Economy") %>%
  summarise(Passengers = sum(Passengers)) %>%
  features(Passengers, features = guerrero) %>%
  pull(lambda_guerrero)

ansett_boxcox <- ansett_na %>%
  filter(Airports == "MEL-SYD", Class == "Economy") %>%
  as_tsibble(index = Week) %>%
  mutate(box_cox = box_cox(Passengers, lambda = ansett_la))

autoplot(ansett_boxcox, box_cox) +
  labs(title = "After Box-Cox",
       x = "Week & Year",
       y = "Transformed Passengers")
```

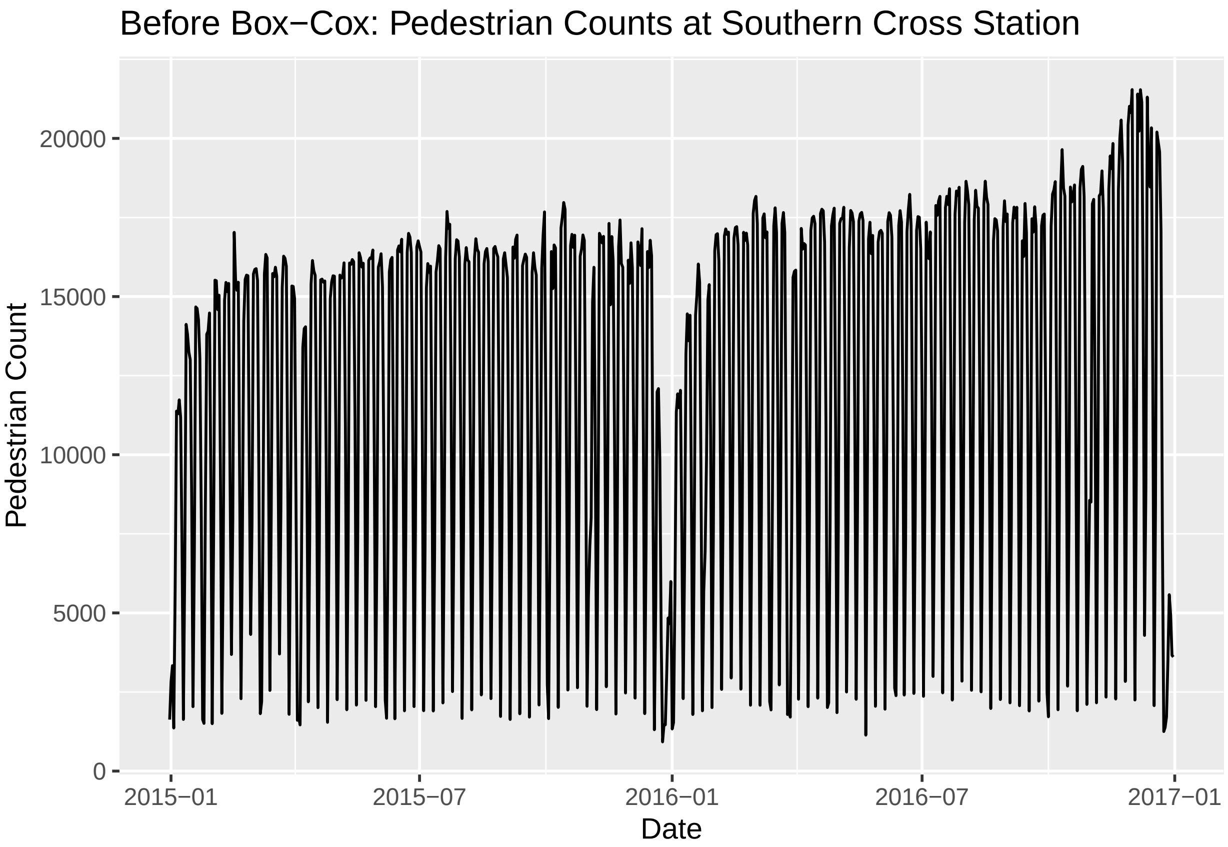


```r
autoplot(ansett_boxcox, box_cox) +
  labs(title = "After Box-Cox",
       x = "Week & Year",
       y = "Transformed Passengers")
```

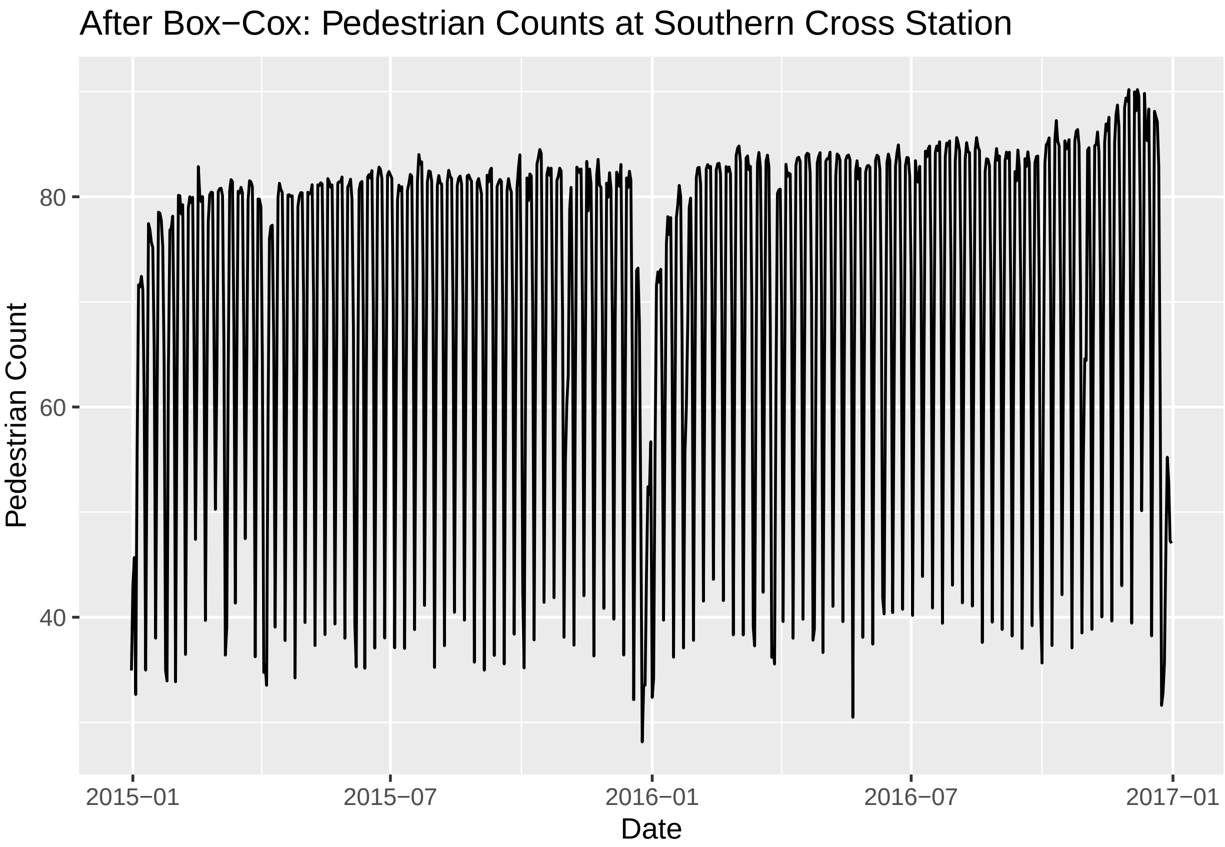


Pedestrian counts at Southern Cross Station from pedestrian

```
pedestrian_na <- pedestrian %>%
  drop_na(Count)

ped_day <- pedestrian_na %>%
  filter(Sensor == "Southern Cross Station") %>%
  index_by(Date = as.Date(Date_Time)) %>%
  summarise(Count = sum(Count))

autoplot(ped_day, Count) +
  labs(title = "Before Box-Cox: Pedestrian Counts at Southern Cross Station",
       x = "Date",
       y = "Pedestrian Count")
```



```r
pedestrian_la <- ped_day %>%
  summarise(Count = sum(Count)) %>%
  features(Count, features = guerrero) %>%
  pull(lambda_guerrero)

pedestrian_boxcox <- ped_day %>%
  mutate(box_cox = box_cox(Count, lambda = pedestrian_la))

autoplot(pedestrian_boxcox, box_cox) +
  labs(title = "After Box-Cox: Pedestrian Counts at Southern Cross Station",
       x = "Date",
       y = "Pedestrian Count")
```



3.7 - Review of Five years of Gas Data from aus_production

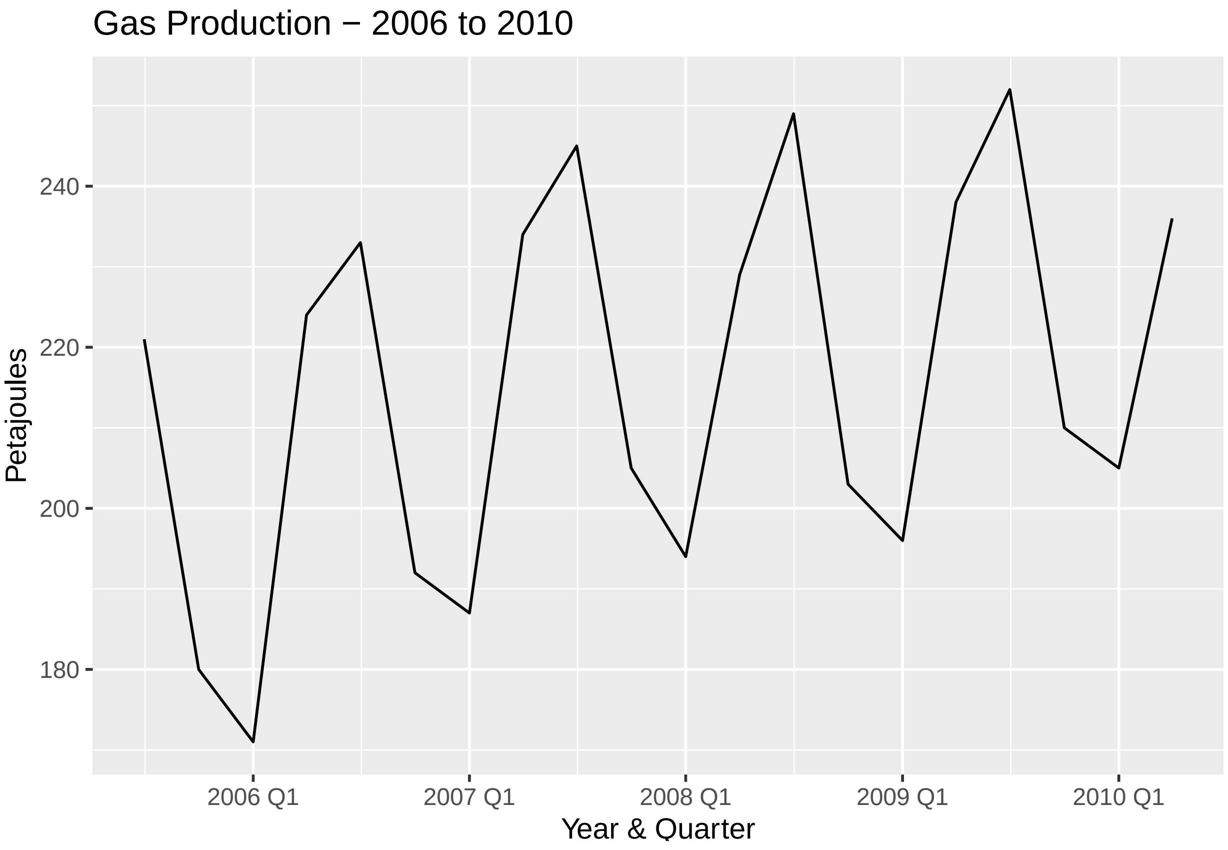
This question asks me to consider the last five years of Gas data from aus_production. The gas variable assignment is code provided by the exercise.

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

Gas production has seasonality that sends it up over the course of Q2, peaks in Q3, and then goes down in Q4. The bottom is in Q1. The peaks have been graduaally increasing over time.

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

autoplot(gas) +
  labs(title = "Gas Production - 2006 to 2010",
               x = "Year & Quarter",
               y = "Petajoules")
```

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



Classical Decomposition and Interpretation

I will answer 7 B and C together.

B. Use classical_decomposition with type=multiplicative to calculate the trend-cycle and seasonal indices

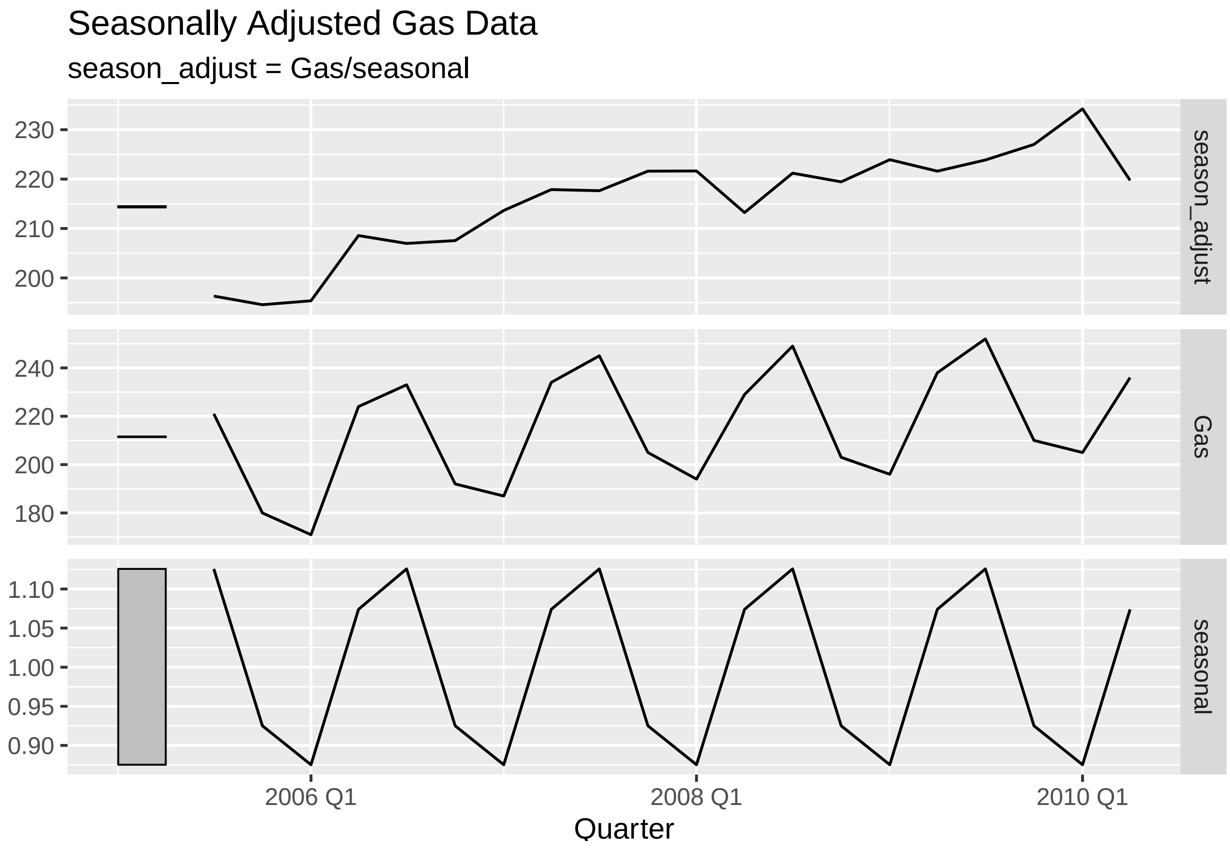
```r
gas_decomp <- gas |>
  as_tsibble() |>
  model(classical_decomposition(Gas, type = "multiplicative")) |>
  components()
```

C. Do the results support the graphical interpretation from part A?

- The results support my interpretation from part A because the general trend remains upwards and the seasonality remains consistent. Charts are all about scale and, relative to 100 to 300 Petajoules a quarter for the trend and seasonal charts, the .9 to 1.1 Petajoule for random doesn't really mean much.

D. Compute and plot seasonally adjusted data

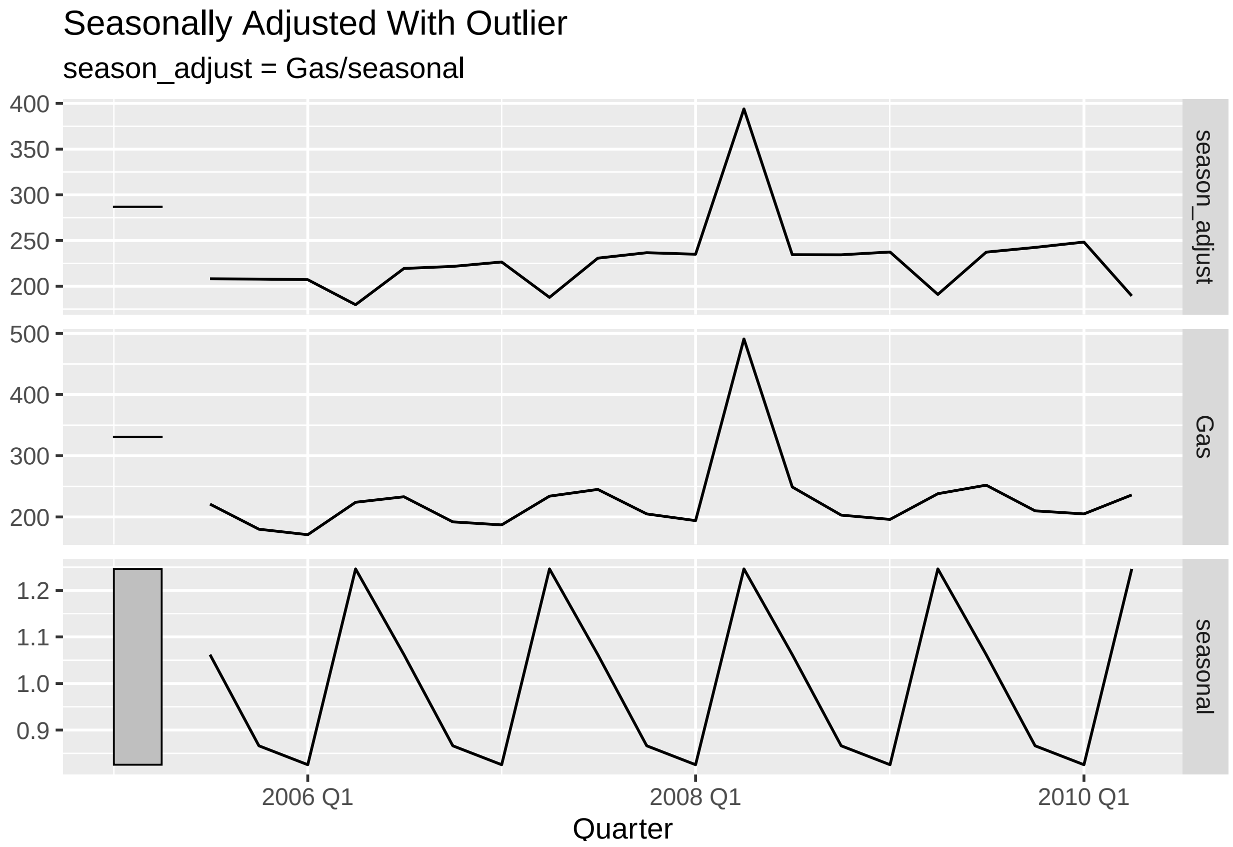
```r
autoplot(gas_decomp, season_adjust) +
  ggtitle("Seasonally Adjusted Gas Data")
```



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?

The outlier causes stable trends for both the seasonal adjustment and standard trend line to essentially become flat, with a huge spike in the middle. The standard seasonal adjustment shifts a bit but remains stable. The general effect is to warp the data, as outliers do, and cause distortions.

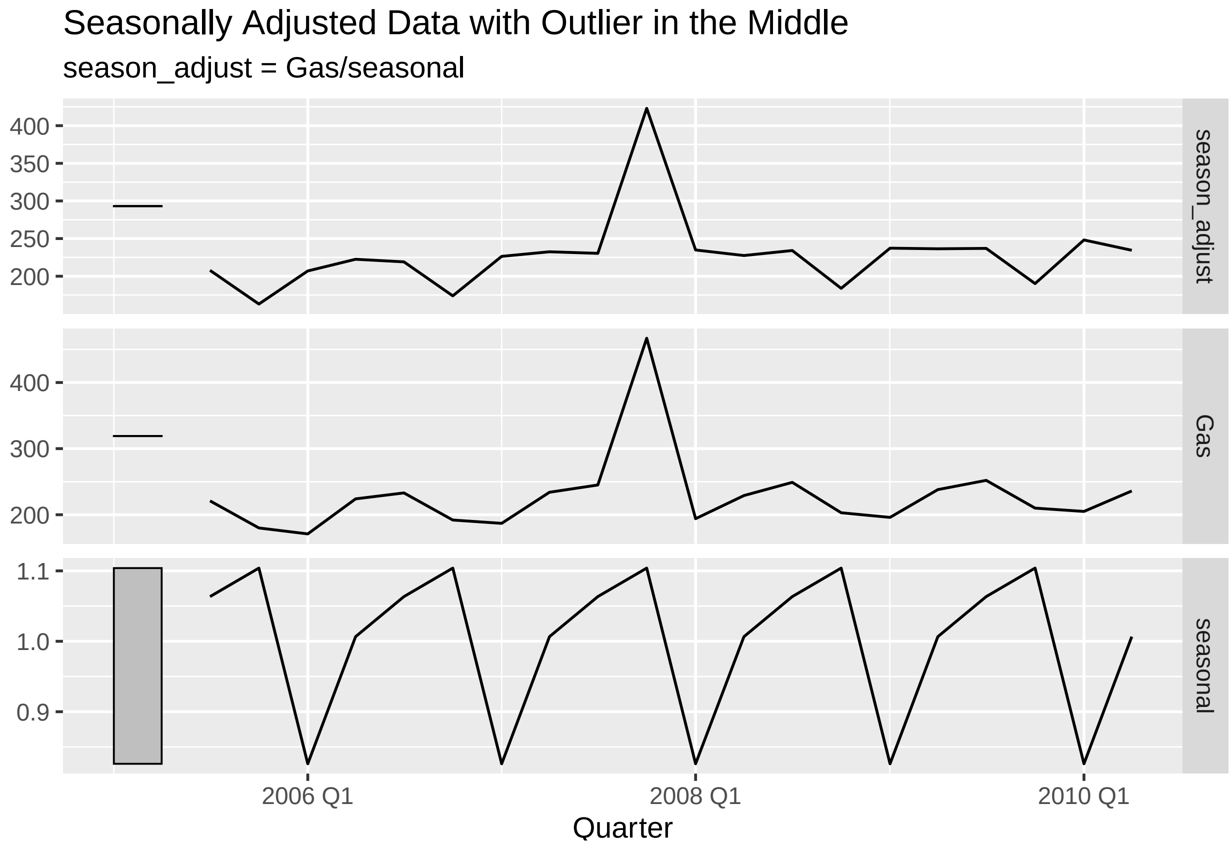
```r
gas_outlier <- gas |>
  mutate(Gas = if_else(row_number() == 12, Gas + 262, Gas))

gas_decomp_outlier <- gas_outlier |>
  as_tsibble() |>
  model(classical_decomposition(Gas, type = "multiplicative")) |>
  components()

autoplot(gas_decomp_outlier, season_adjust) +
  ggtitle("Seasonally Adjusted With Outlier")
```



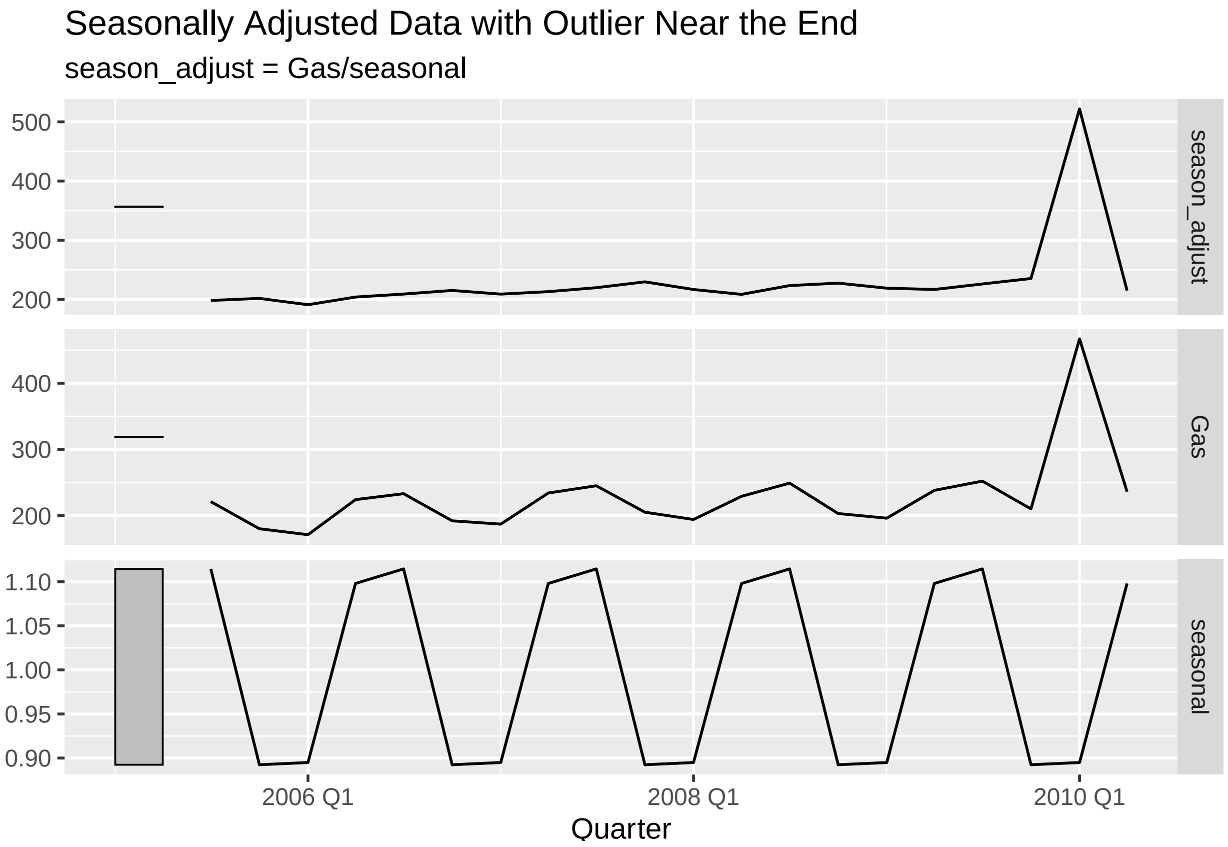
F. Does it make any difference if the outlier is near the end rather than in the middle of the time series?

Yes. The outlier in the middle causes the seasonal adjustment trend to become disturbed and even, with the spike in the middle. When it's at the end, this line remains steady and then spikes. This means an outlier at the end is an anomoly but an outlier in the middle disturbs the whole trend.

```
nrows <- nrow(gas)
gas_outlier_middle <- gas |>
  mutate(Gas = if_else(row_number() == round(nrows / 2), Gas + 262, Gas))

gas_outlier_end <- gas |>
  mutate(Gas = if_else(row_number() == nrows - 1, Gas + 262, Gas))

gas_decomp_outlier_middle <- gas_outlier_middle |>
  as_tsibble() |>
  model(classical_decomposition(Gas, type = "multiplicative")) |>
  components()


gas_decomp_outlier_end <- gas_outlier_end |>
  as_tsibble() |>
  model(classical_decomposition(Gas, type = "multiplicative")) |>
  components()


autoplot(gas_decomp_outlier_middle, season_adjust) +
  ggtitle("Seasonally Adjusted Data with Outlier in the Middle")
```



```
autoplot(gas_decomp_outlier_end, season_adjust) +
  ggtitle("Seasonally Adjusted Data with Outlier Near the End")
```

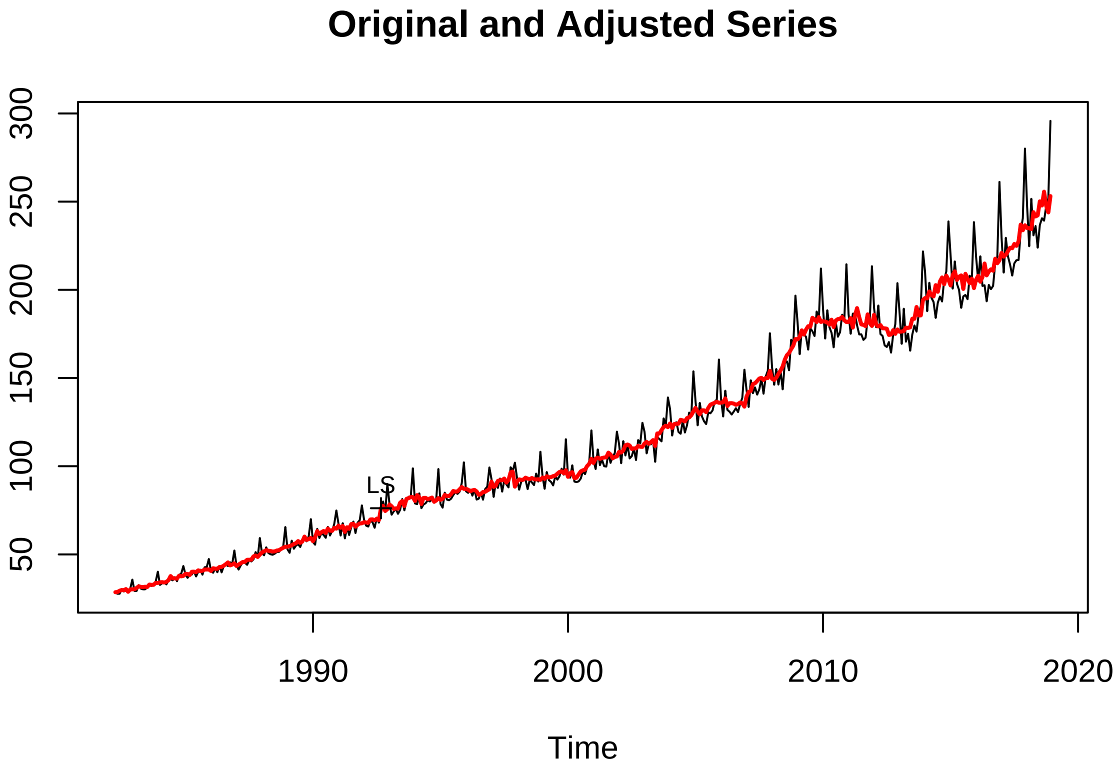


3.8 Decomposition Using X-11 of aus_retail

This exercise asks me to decompose the aus_retail dataset using the X-11 method and provide commentary on outliers or unusual features. X-11 is a techique developed by the American and Canadian governments that breaks data down into its different components.

What's interesting is that the adjusted line appears to be a microscopic version of the original line. It's not perfeftly smooth or straight, suggesting that that the underlying turberlence is real and structural. Everything has been growing proportionally, more or less. I'm not sure if exponential is the right word but, just by following the spikes all the way up the chart, you can see them get larger and larger.

```
nsw_retail <- aus_retail %>%
  filter(State == "Tasmania", Industry == "Food retailing") %>%
  select(Month, Turnover)

nsw_retail_ts <- ts(nsw_retail$Turnover, start = c(1982, 4), frequency = 12)
x11_decomp <- seas(nsw_retail_ts, x11 = "")
plot(x11_decomp)
```



3.9 Australian Civilian Labor Force Review

The answer to this question assumes you have access to figures 3.19 and 3.20 from the book exercise and can review them as you read these answers.

A. The results of the decompositions show that the civilian labor force has seen very solid growth. The seasonal component is on a scale that's order of magnitudes smaller than the value or trend charts, which means it's not generally consequential to the overall dataset. It's interesting to see that the trend line is not negatively impacted at any time, I would have expected something at some point.

B. The 1991-1992 recession is visible to an extent, but scales also matter here. There's a visibly dramatic drop on the remainder chart but it only sits on a scale of hundreds while the others are on thousands. The value trend was impacted and didn't return to the course it would have been on without it until the end of the series in 1995.