Name: Your name here
Due: 2024/09/30

Homework 3

Be sure to submit **both** the .pdf and .qmd file to Canvas by Monday, September 30th at 11:59 pm.

0. [1 pt] With whom did you work on this assignment?

1. [11 pt] We will focus on a data set describing weekly avocado sales volume and price in the United States between 2015 and 2018 for this question.

   a) [2 pt] Read the data in (naming it `avocado`), filter to sales of conventional avocados in Las Vegas, and create a new column, called `volume1000`, that represents the total volume of sales in 1000s.

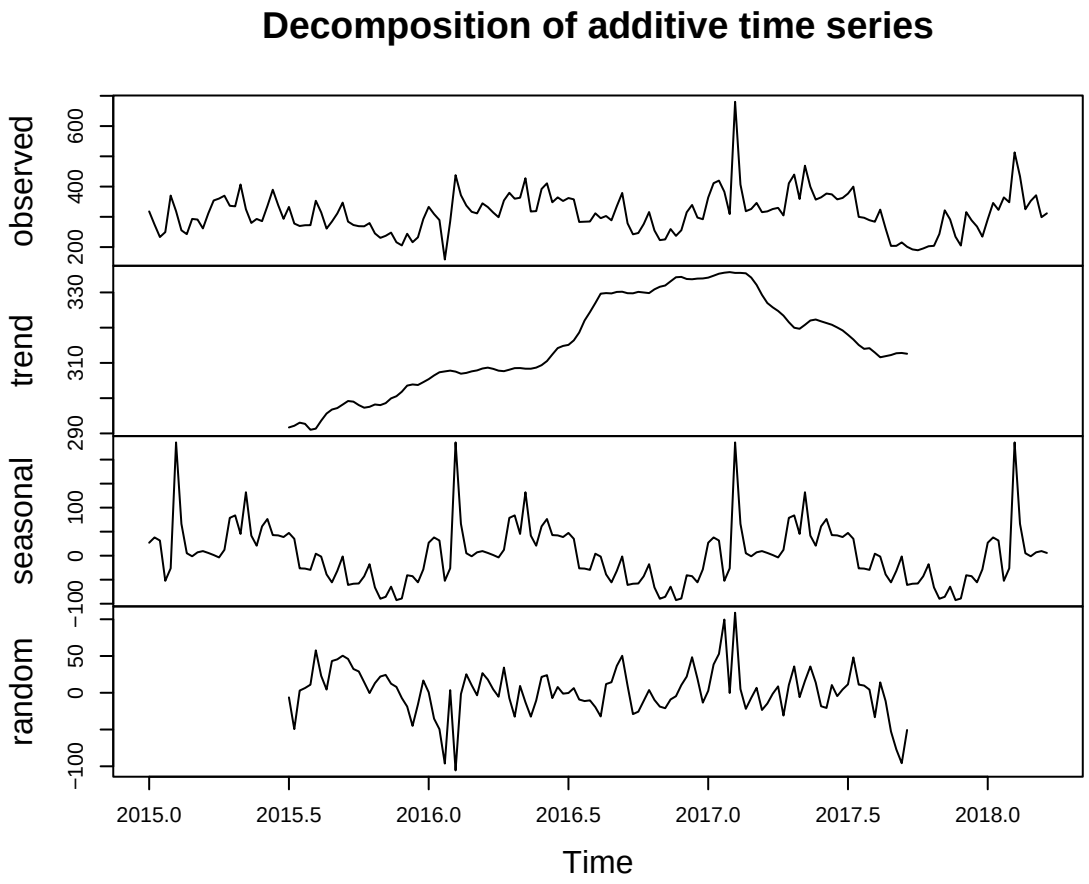
```
# load the data, sort by date and filter to conventional sales in Albany
avocado <- readr::read_csv("avocado.csv") %>%
  arrange(Date) %>%
  filter(
    region == "LasVegas",
    type == "conventional"
  ) %>%
  mutate(volume1000 = `Total Volume`/1000)
```

   b) [3 pt] Create a `ts` object with the `volume1000` vector (called `avo_ts`), create an additive decomposition of that time series (called `avo_decomp`), and plot that decomposition.

> **Tip**
>
> Be sure to pay attention to how the data set is arranged with respect to date.

```
# construct weekly ts
avo_ts <- ts(
  avocado$volume1000,
  start = with(avocado, c(year(Date[1]), week(Date[1]))),
  end = with(
    avocado,
    c(year(Date[nrow(avocado)]), week(Date[nrow(avocado)]))
  ),
  freq = 52
)
```

```
# decomposition
avo_decomp <- decompose(avo_ts, "additive")
plot(avo_decomp)
```



c) [1 pt] Describe the time series decomposition in terms of its trend and seasonal components.

There is clear evidence of a positive trend and strong seasonal elements. Over time, the sale of avocados increases fairly substantially. On top of that, there is a seasonal trend in which the sales spike early in the year, then again in the summer but to a lesser extent, then fall off through the winter.

d) [1 pt] You should notice that the trend and random components are significantly (in the english way, not the statistics way) shorter than the other two series. Why is this?

Half of the frequency is discarded on each of the time series to calculate the moving average, which is 26 observations on each side - 52 total which is nearly 1/3 of the data!

e) [1 pt] You should see a rather large spike in avocado sales during the beginning of each year in the decomposition. Hypothesize an explanation for this spike.

> Super bowl! Classic USA.

f) [1 pt] The code below attempts to create a data frame of the random component over time, but R kicks back an error. Investigate the source of this error, and describe what is causing the problem.

```
res_tbl <- tibble(
  dt = avocado$Date,
  res = avo_decomp$random
)
```

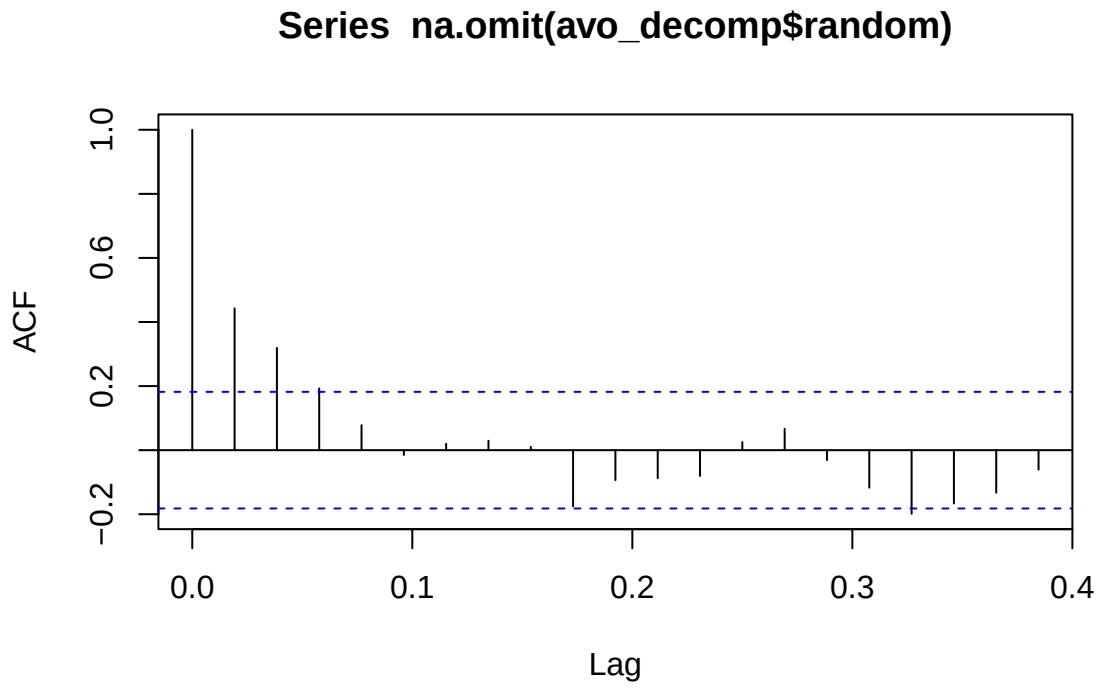
> The year 2017 has 53 weeks! Well, 53 Sundays anyway. This is another (hopefully) valuable lesson in exploring your data prior to analysis! In this case, one fix might be to remove the week that begins on 2017-12-31 from the data, since it contains only one day from the year 2017 (which would once again balance the number of weeks in each year).

```
avocado %>% group_by(year) %>% summarize(count = n())
```

```
# A tibble: 4 x 2
   year count
  <dbl> <int>
1  2015    52
2  2016    52
3  2017    53
4  2018    12
```

g) [2 pt] We are going to ignore the problem for now (:]). Provide a correlogram of the residual error series and comment on whether there appears to be leftover serial correlation.

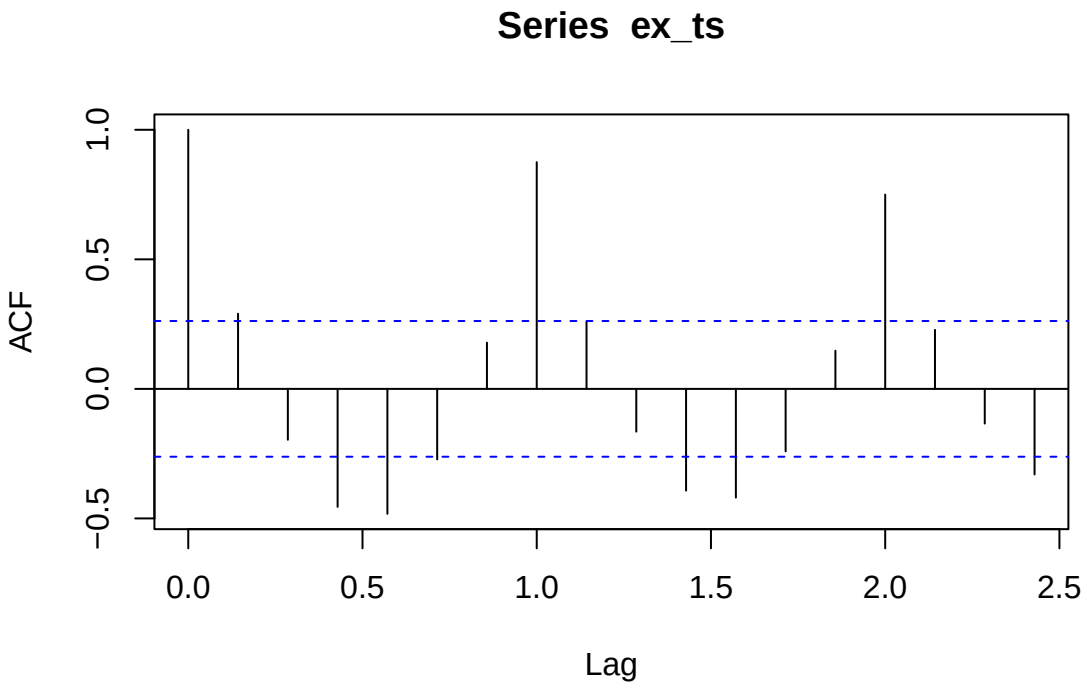
```
acf(na.omit(avo_decomp$random))
```



> The sample autocorrelation at lag 1 is nearly 0.5, which suggests that there is definitely some leftover serial correlation that we should address. To be continued…

2. [1 pt] The code below creates a time series out of the same seven values on repeat, and generates a correlogram for that series. Describe what is happening in the correlogram at time lags that are a multiple of seven. Why should that make sense?

```
ex_ts <- ts(
  rep(-3:3, 8),
  start = c(1,1),
  end = c(8, 7),
  freq = 7
)
acf(ex_ts)
```



Observations that are seven time points apart are *identical*, so the correlation is extremely close to 1. This is one way to diagnose unaccounted for seasonal variability! You would expect to see large spikes in the ACF plot corresponding to the seasonal frequency.

3. [6 pt] The rest of the questions on this assignment are fairly math-heavy, but in order to understand time series analysis, it is essential to understand the concepts of stationarity, covariance, and correlation. If the following questions feel hard, *that is okay* - it does not mean you cannot learn time series. I encourage you to ask for help!

   Note: When I use the word *prove* below, I do not mean in a rigorous mathematical sense (though, by all means give it a shot!). I only ask that you provide logic and sound mathematical reasoning (meaning I should see an equation or two in each response).

   a) [3 pt] *Prove* that a model cannot be second-order stationary if the model is not stationary in the mean.

   Hint

   What do we know about the autocovariance function, $\gamma(k)$ for a second-order stationary process?

   Suppose that a model is *not* stationary in the mean. This implies that the mean function, $\mu(t)$ is non-constant and is instead a function of t. That is, $\mu(t) = \mu_t$. The autoco-

variance function for this model that is not stationary in the mean would therefore be

$$\gamma(k,t) = E\left[(X_t - \mu_t)(X_{t+k} - \mu_t)\right]$$

For a second-order stationary process, we know that the acvf must depend only on the time lag between observations, and not on the particular value of t. Therefore, a process that is not stationary in the mean cannot be second-order stationary.

b) [3 pt] Recall that the population autocorrelation function is given by $\rho_k = \frac{\gamma_k}{\sigma^2}$. *Prove* that ρ_0 must always equal 1.

Short and sweet on this one - it follows immediately from the definition.

$$\rho_0 = \frac{\gamma_0}{\sigma^2} = \frac{E\left[(X_t - \mu)(X_{t+0} - \mu)\right]}{E[(X_t - \mu)^2]} = \frac{E\left[(X_t - \mu)(X_t - \mu)\right]}{E[(X_t - \mu)^2]} = \frac{E\left[(X_t - \mu)^2\right]}{E[(X_t - \mu)^2]} = 1$$

4. [4 pt] (This question is admittedly a nasty piece of work, but valuable to understand). Recall that the sample autocorrelation function, r_k, is defined as

$$r_k = \frac{c_k}{c_0}$$

but the population autocorrelation function, ρ_k is defined as

$$\rho_k = \frac{\gamma_k}{\sigma^2}.$$

Why do we define $r_k = \frac{c_k}{c_0}$, rather than $r_k = \frac{c_k}{s^2}$, which is arguably more natural? Support your answer with something proof-ish.

The answer to this is actually delectably simple. First, recall (from 2b) that the autocorrelation at lag 0 *must* be exactly 1, since the correlation between a sequence of numbers and itself must be 1. That is all we need to know to reason through this question.

I provide some mathematical details below, but the idea is that in order to guarantee that $r_0 = 1$, we cannot use the sample variance s^2, and must instead use c_0.

Suppose that we do use the sample variance. That is, let $r_k = \frac{c_k}{s^2}$. Now consider calculating r_0.

$$r_0 = \frac{c_0}{s^2} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(x_t - \bar{x})}{\frac{1}{n-1}\sum_{t=1}^{n}(x_t - \bar{x})^2} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}{\frac{1}{n-1}\sum_{t=1}^{n}(x_t - \bar{x})^2} = \frac{n-1}{n} \neq 1.$$

Uh oh! We know that r_0 must be 1, so the sample variance won't work. It turns out that c_0 is exactly the sample variance, except you divide by n rather than $n-1$. And of course, it is trivial to show that $r_0 = 1$ if we use the proper denominator, since $r_0 = \frac{c_0}{c_0} = 1$.