Name: Your name here
Due: 2024/10/07

Homework 4

Be sure to submit **both** the .pdf and .qmd file to Canvas by Monday, October 7th at 11:59 pm.

0. [5 pt] Please complete this Google Form to provide feedback on how the semester is going so far. The form is anonymous, so I cannot verify that you have actually completed the form. Please do it :(It is for your benefit as well as my own!

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

2. [17 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) [1 pt] Read the data in (naming it `avocado`) and filter to sales of conventional avocados in `Albany`.

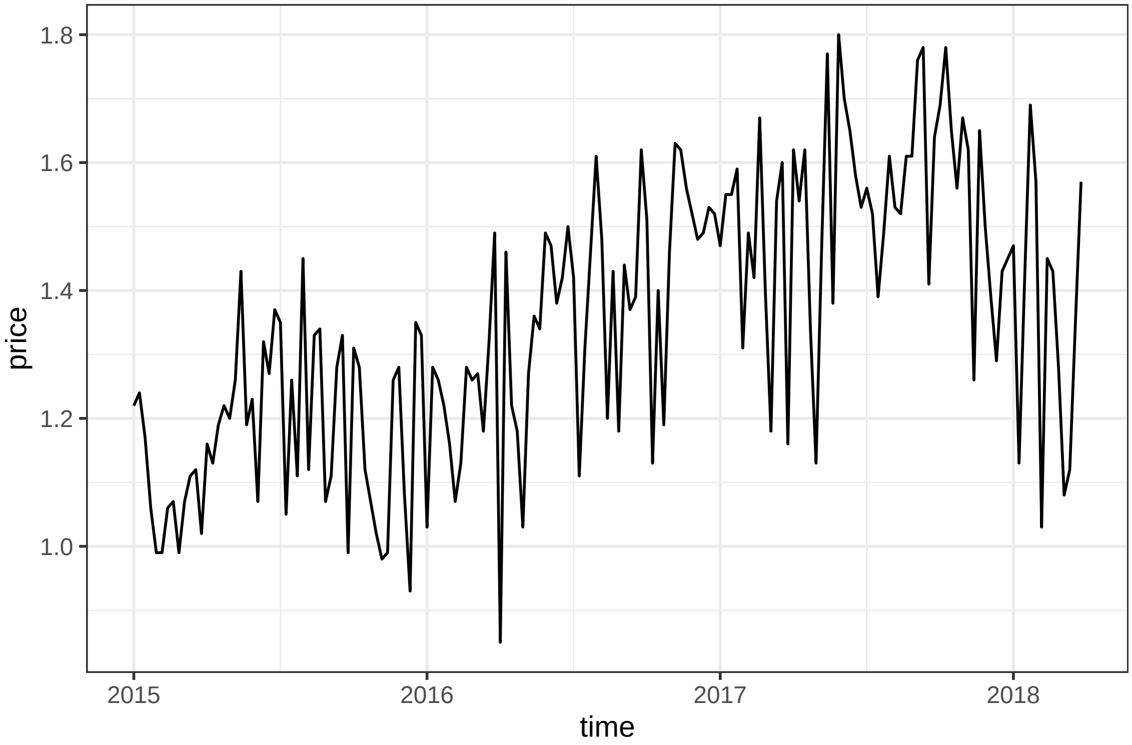
   ```r
   # load the data, sort by date and filter to conventional sales in Albany
   avocado <- readr::read_csv("avocado.csv") %>%
     arrange(Date) %>%
     filter(
       region == "Albany",
       type == "conventional"
     )
   ```

   b) [2 pt] Create a `ts` object with the `AveragePrice` vector (called `avo_ts`) and plot the series.

   > **Tip**
   >
   > Be sure to pay attention to how the data set is arranged with respect to date. Additionally, **do not** specify an end date. This is one way to handle the fact that there are 53 Sundays in 2017 (omitting an end date forces R to treat the week that begins on 12/31/2017 as the first week of 2018).

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

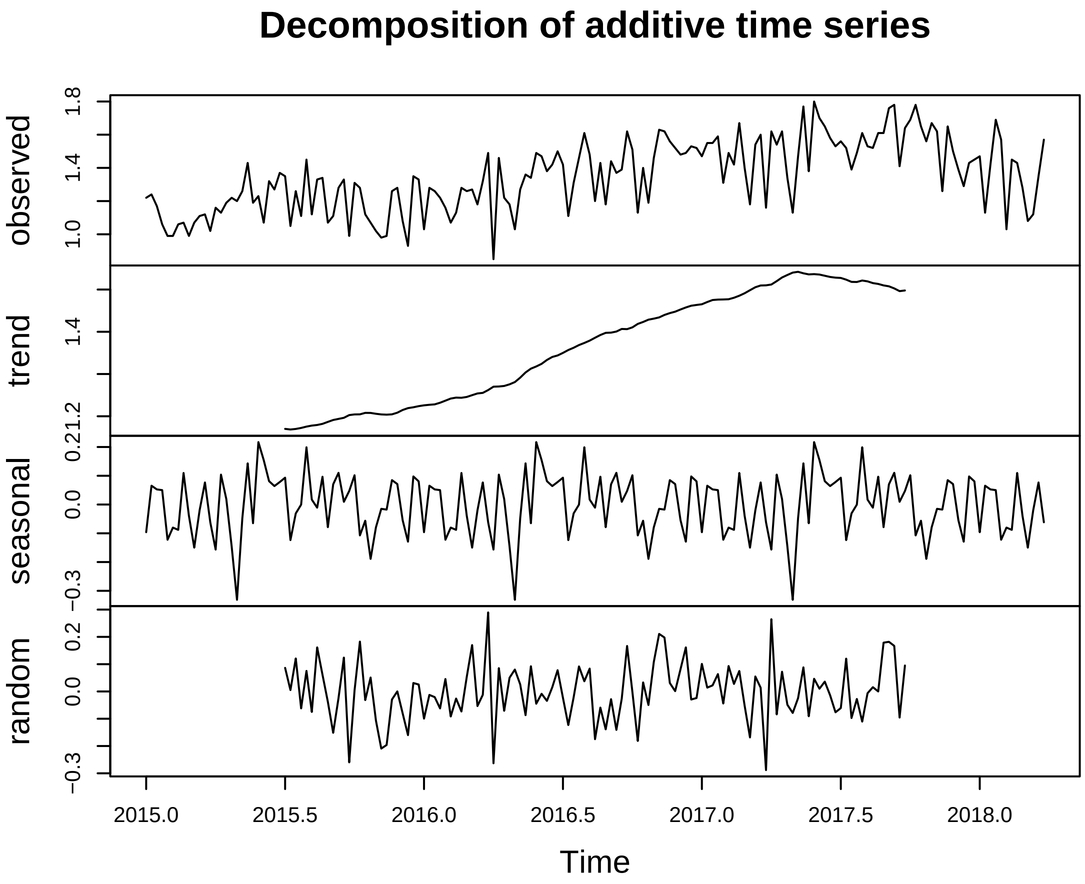
```
# plot
tibble(
  price = avo_ts,
  time = time(avo_ts)
) %>%
  ggplot(aes(x = time, y = price)) +
  geom_line() +
  theme_bw()
```



c) [2 pt] Describe the series in terms of trend and seasonality. Decomposing the series might help.

> There is evidence of a positive trend over time. The seasonal term is not as clear, but perhaps there is some evidence of a drop in avocado prices in the winter.

```
# construct weekly ts
plot(decompose(avo_ts))
```



d) [1 pt] Create a reduced version of the `avo_ts` time series, called `avo_ts_red`, that only spans 2015 to 2017.

```
avo_ts_red <- window(
  avo_ts,
  start = c(2015, 1),
  end = c(2017, 52)
)
```

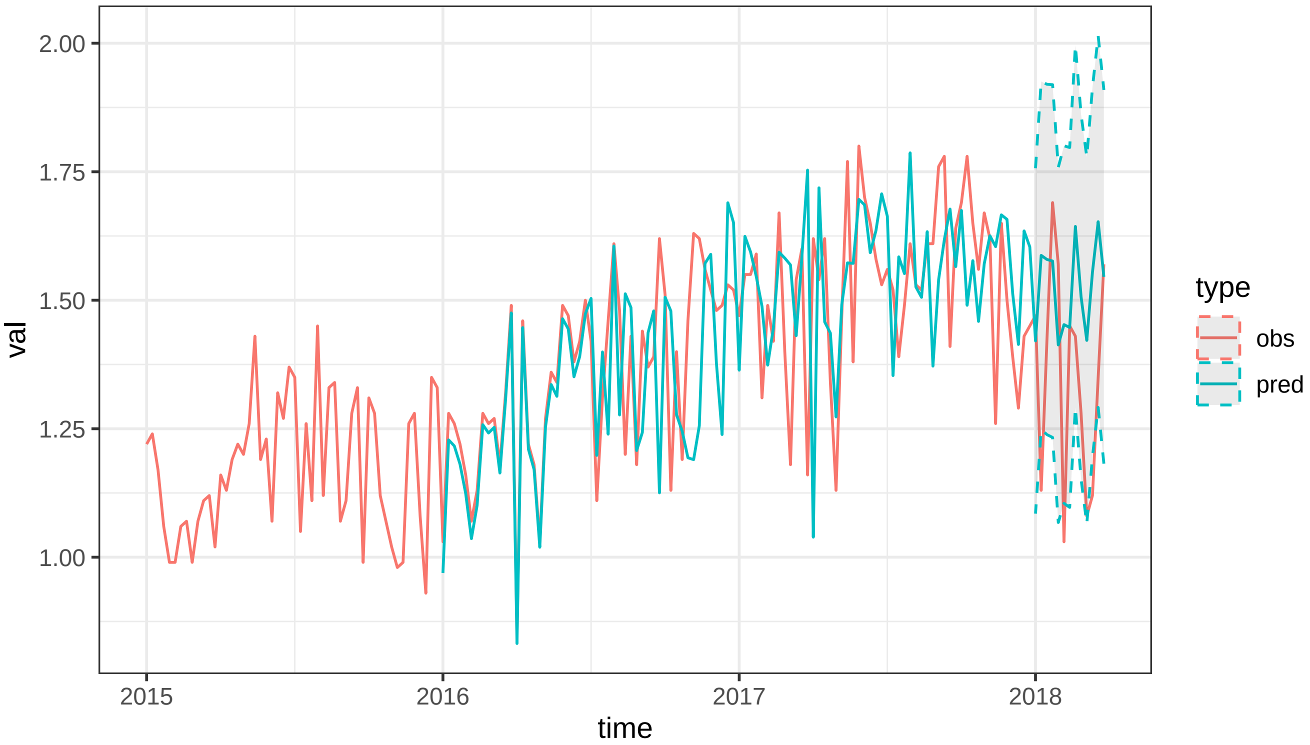
e) [1 pt] Fit an additive Holt-Winters model, called `avo_hw1`, to the reduced time series and allow R to estimate the smoothing parameters.

```
avo_hw1 <- HoltWinters(avo_ts_red)
```

f) [4 pt] Create an object, called `avo_hw1_pred`, that predicts the first 13 weeks of 2018 and include prediction intervals. Plot the original time series, fitted series, and forecasted series on the same plot. How well does the forecast predict the avocado prices from the complete time series?

```
# prediction
avo_hw1_pred <- predict(
  avo_hw1,
  n.ahead = 13,
  prediction.interval = TRUE
)

bind_rows(
  tibble(
    time = time(avo_ts),
    val = c(avo_ts)
  ) %>% mutate(type = "obs"),
  bind_rows(
    tibble(
      time = time(avo_hw1$fitted[,1]),
      val = c(avo_hw1$fitted[,1])
    ),
    tibble(
      time = time(avo_hw1_pred),
      val = c(avo_hw1_pred[,1]),
      lwr = c(avo_hw1_pred[,3]),
      upr = c(avo_hw1_pred[,2])
    )
  )  %>% mutate(type = "pred")
) %>%
  ggplot(aes(x = time, y = val, col = type)) +
  geom_line() +
  geom_ribbon(
    aes(ymin = lwr, ymax = upr),
    linetype = 2,
    alpha = 0.1
  ) +
  theme_bw()
```



> It honestly does a pretty poor job, significantly overestimating the true time series.

g) [1 pt] Calculate the sum of squared differences between the forecasted time series and true value of the time series in 2018 (and print that value).

```
diffs <- window(
  avo_ts,
  start = c(2018, 1)
) - avo_hw1_pred[,1]
sum(diffs^2)
```

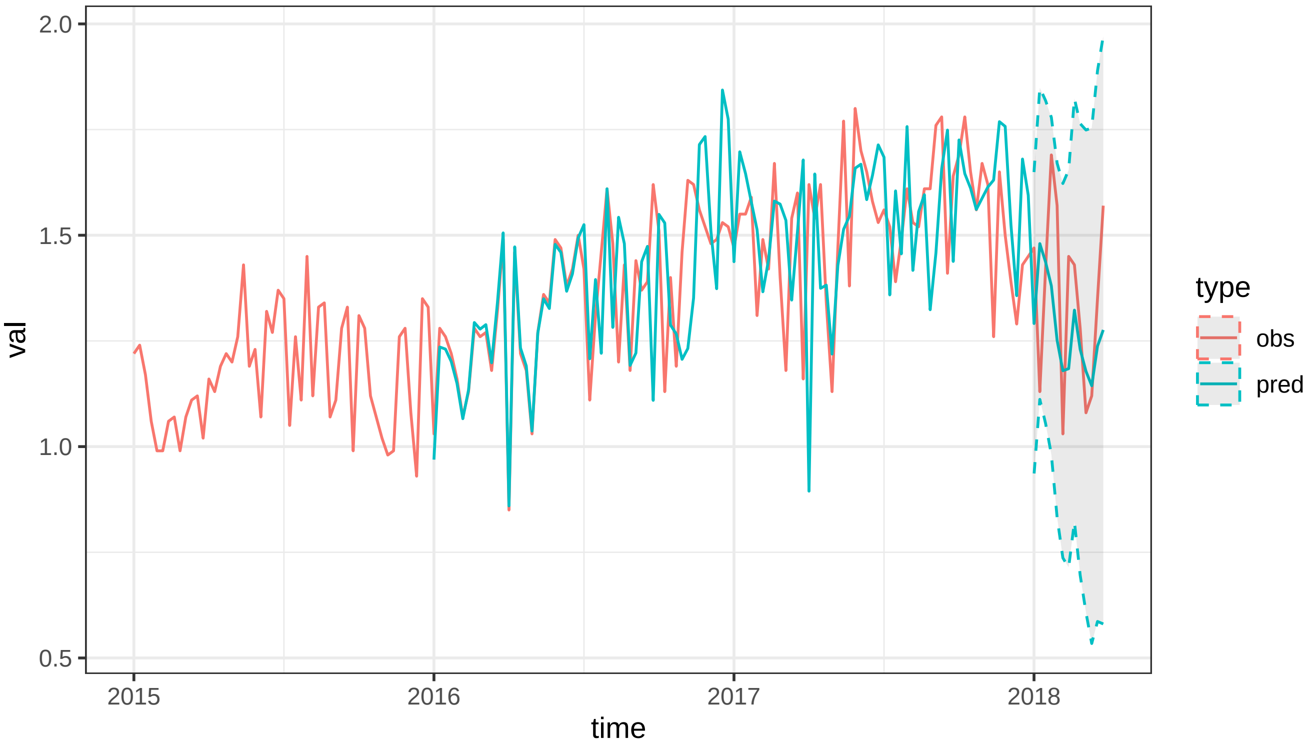
```
[1] 0.9481649
```

h) [5 pt] Find values for the smoothing parameters that result in better forecasts of avocado prices in 2018. Recreate the figure from part f for the new model, and print out the sum of squared differences between the forecasted and observed time series for the new model. What do you notice about the variability of the forecasted prices for the new model relative to the old model?

```
# new model
avo_hw2 <- HoltWinters(
  avo_ts_red,
  alpha = .2,
  beta = .2,
  gamma = .2
)

# prediction
avo_hw2_pred <- predict(
  avo_hw2,
  n.ahead = 13,
  prediction.interval = TRUE
)

# plot
bind_rows(
  tibble(
    time = time(avo_ts),
    val = c(avo_ts)
  ) %>% mutate(type = "obs"),
  bind_rows(
    tibble(
      time = time(avo_hw2$fitted[,1]),
      val = c(avo_hw2$fitted[,1])
    ),
    tibble(
      time = time(avo_hw2_pred),
      val = c(avo_hw2_pred[,1]),
      lwr = c(avo_hw2_pred[,3]),
      upr = c(avo_hw2_pred[,2])
    )
  )  %>% mutate(type = "pred")
) %>%
  ggplot(aes(x = time, y = val, col = type)) +
  geom_line() +
  geom_ribbon(
    aes(ymin = lwr, ymax = upr),
    linetype = 2,
    alpha = 0.1
  ) +
  theme_bw()
```



```r
# SSFE
diffs <- window(
  avo_ts,
  start = c(2018, 1)
) - avo_hw2_pred[,1]
sum(diffs^2)
```

```
[1] 0.5685707
```

To obtain better forecasts, we had to increase the smoothing parameter associated with the level. As a result, the smoother places more weight on the recent observations, resulting in quickly compounding error. Therefore, the variability is greater! As more weight is placed on recent observations, the variability associated with our forecasts increases.

3. [6 pt] In class, we saw that the estimate of the non-stationary mean of the exponential smoother is

$$a_t = \alpha x_t + (1-\alpha)a_{t-1}$$

for $0 < \alpha < 1$. This formula is defined *recursively*, meaning that each term is a function of the previous term. We will encounter many time series models that are defined recursively, so it is helpful to fully understand what it is implied by this kind of model.

a) [1 pt] Show that a_t may also be defined as $a_t = \alpha(x_t - a_{t-1}) + a_{t-1}$.

Just a bit of algebra for this one.

$$\begin{aligned} a_t &= \alpha x_t + (1-\alpha)a_{t-1} \\ &= \alpha x_t + a_{t-1} - \alpha a_{t-1} \\ &= \alpha(x_t - a_{t-1}) + a_{t-1} \end{aligned}$$

b) [4 pt] Show that $a_t = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \ldots$.

We just need to back-substitute the formula. To make our lives easier, let us establish some results first.

$$a_t = \alpha(x_t - a_{t-1}) + a_{t-1} = \alpha x_t - \alpha a_{t-1} + a_{t-1} = \alpha x_t + (1-\alpha)a_{t-1}$$

Similarly,

$$\begin{aligned} a_{t-1} &= \alpha x_{t-1} + (1-\alpha)a_{t-2} \\ a_{t-2} &= \alpha x_{t-2} + (1-\alpha)a_{t-3} \end{aligned}$$

That is all we need to prove the result! Now we just start backwards substituting.

$$\begin{aligned} a_t &= \alpha x_t + (1-\alpha)a_{t-1} \\ &= \alpha x_t + (1-\alpha)\left[\alpha x_{t-1} + (1-\alpha)a_{t-2}\right] \\ &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + (1-\alpha)^2\left[\alpha x_{t-2} + (1-\alpha)a_{t-3}\right] \\ &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + (1-\alpha)^3 a_{t-3} \\ &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \ldots \end{aligned}$$

c) [1 pt] Recall that $0 < \alpha < 1$. Comment on what the result in part b implies about how the weight associated with recent observations changes as we move backwards in time.

It decays as we move backwards in time!

4. [5 pt] Show that

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$$

.

You may use any (read: all) of the following results in your proof:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

$$\sum_{i=1}^{n} X_i \sum_{j=1}^{m} Y_j = \sum_{i=1}^{n}\sum_{j=1}^{m} X_i Y_j$$

Just need to plug in the results that we are given.

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = E\left(\sum_{i=1}^{n} X_i \sum_{j=1}^{m} Y_j\right) - E\left(\sum_{i=1}^{n} X_i\right) E\left(\sum_{j=1}^{m} Y_j\right)$$

by the first result. By the third result,

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = E\left(\sum_{i=1}^{n}\sum_{j=1}^{m} X_i Y_j\right) - E\left(\sum_{i=1}^{n} X_i\right) E\left(\sum_{j=1}^{m} Y_j\right)$$

Next, we distribute the expectations inside the sums by the second result.

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} E(X_i Y_j) - \sum_{i=1}^{n} E(X_i) \sum_{j=1}^{m} E(Y_j)$$

We next apply the third result again,

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} E(X_i Y_j) - \sum_{i=1}^{n}\sum_{j=1}^{m} E(X_i)E(Y_j)$$

A bit of algebra, since the sums span the same index,

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \left[E(X_i Y_j) - E(X_i)E(Y_j)\right]$$

Finally, we apply the first result again,

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_i)$$

And there you have it - a bona fide proof!