Assignment 1.1: Module 1 Exercises

Zachariah Freitas

2022-10-30

Chapter 1:

Approaching Forecasting (Page 23) #1-5

Question 1.

Is the goal of this study descriptive or predictive?

Question CH1-Q1 Answer: **Answer:** The goal of this study is to be descriptive. To understand the impact that 9/11 had on long distance passenger travel behavior patterns.

Question 2.

What is the forecast horizon to consider in this task? Are next-month forecasts sufficient?

```
library(readxl)
library(zoo)
```

Question CH1-Q2 Answer:

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(tidyverse)
```

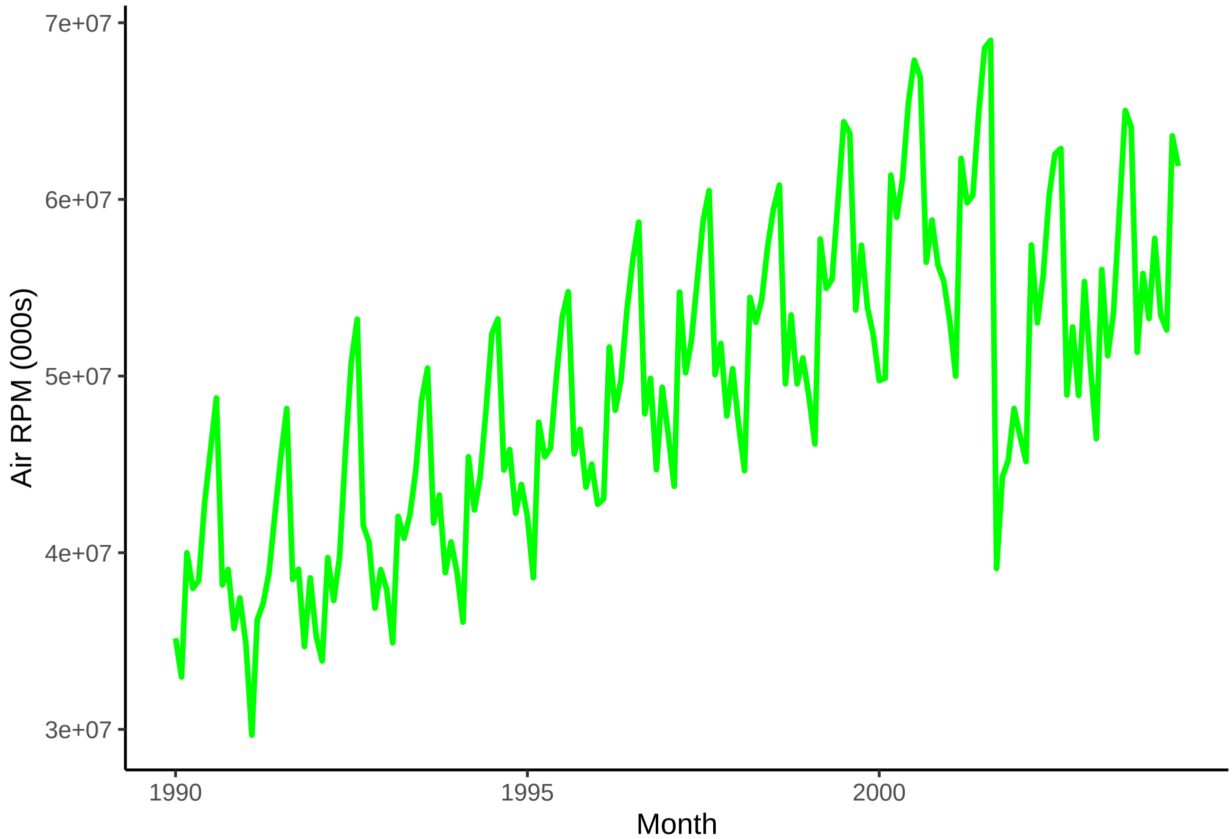
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fpp2) # Plot and Forecast Data
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
## -- Attaching packages ---------------------------------------------- fpp2 2.4 --
## v forecast  8.18     v expsmooth 2.3
## v fma       2.4
```

```
set.seed(506)

Sept11Travel <- read_excel("Sept11Travel.xls",
    col_types = c("date", "numeric", "numeric",
        "numeric"))


ggplot(Sept11Travel, aes(x = Month, y = `Air RPM (000s)`, group = 1)) +
  geom_line(size = 1, color = "green") +
  theme_classic()
```




```
#
# ggplot(Sept11Travel, aes(x = Month, y = `Rail PM`, group = 1)) +
#   geom_line(size = 1, color = "green") +
#   theme_classic()
#
#
# ggplot(Sept11Travel, aes(x = Month, y = `VMT (billions)`, group = 1)) +
#   geom_line(size = 1, color = "green") +
#   theme_classic()
```

Answer: Impacts to 9/11 was felt for years after the attack. You can visually see the impact in the chart above. So looking at a minimum of several months to several years would be appropriate.

Question 3.

What level of automation does this forecasting task require? Consider the four questions related to automation.

Question CH1-Q3 Answer: **Answer:** Given my answers below the level automation required for this task is relatively low.

1. How many series need to be forecasted?

   Answer: Only three, "Air RPM (000s)", "Rail PM", and "VMT (billions)".

2. Is the forecasting an ongoing process or a one time event?

   Answer: This forecast is a one time event. We hope.

3. Which data and software will be available during the forecasting period?

   Answer: A one time pull of Sept11Travel.xls data will be available and R will be used to analyze this data. With the following additional libraries, readxl, zoo, tidyverse, and fpp2.

4. What forecasting expertise will be available **Answer:** Graduate level data scientist will be available for this task.

Question 4.

What does the meaning of t = 1,2,3 in the Air series? Which time period does t = 1 refer to?

Question CH1-Q4 Answer: **Answer:** t = 1,2,3 is an index denoting the time period of interest, with this data t = monthly periods. t = 1 is the fist period in the series, starting in January of 1990.

Question 5.

What are the values for y_1, y_2, and y_3 in the Air series?

Question CH1-Q5 Answer: **Answer:** y_1, y_2, and y_3 in the Air series, is a series of n values measured over n time periods, where y_t denotes the value of the series at time period t. For example, this data is a series of monthly Air RPM (000's), t =1,2,3,... denotes month 1, month 2, and month3; y_1, y_2, and y_3 denote the Air RPM (000's) on months 1, 2, and 3.

Chapter 2:

Time Series Data (Page 43) #3

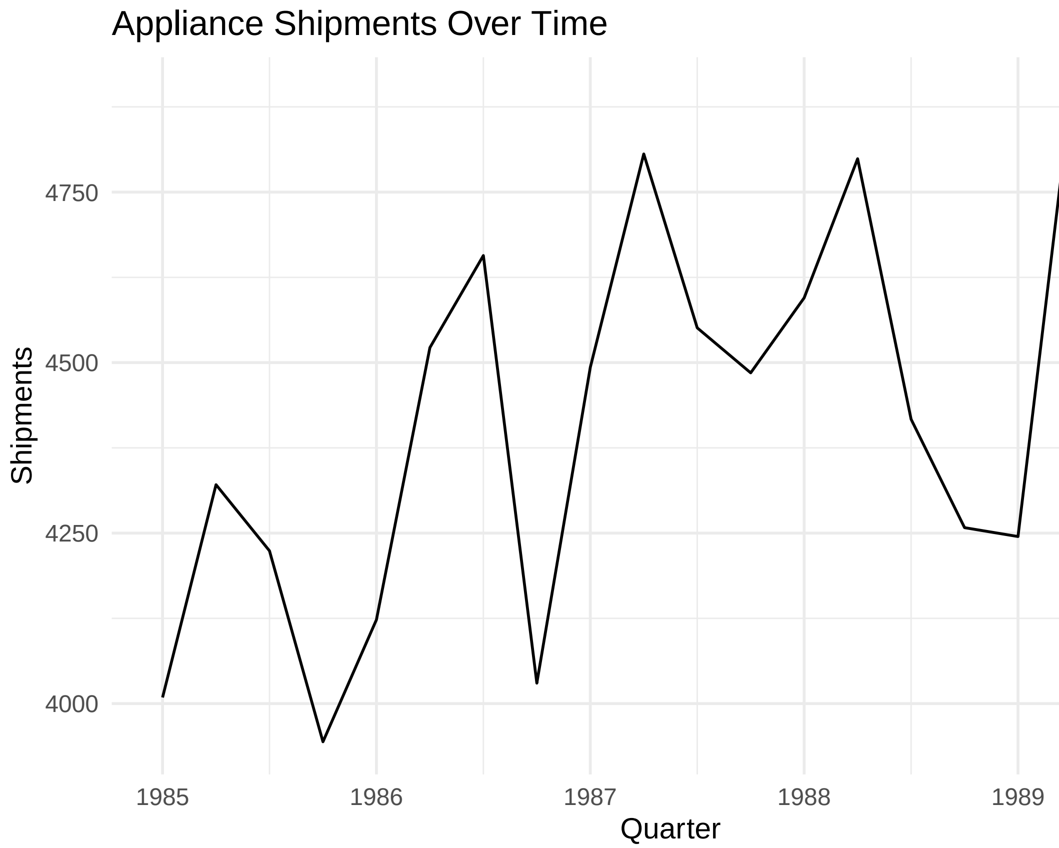
Question 3.

Shipments of Household Appliance: The file ApplianceShipments.xls contains the series of quarterly shipments (in millions of USD) of U.S. household appliances between 1985-1989.

a) Create a well-formatted time plot of the data.

b) Which of the four components (level, trend, seasonality, noise) seem to be present in this series?

```r
library(readxl)
library(zoo)
library(tidyverse)
ApplianceShipments <- read_excel("ApplianceShipments.xlsx")

# Convert Quarter character to date
ApplianceShipments$Quarter <-
  as.Date(
    as.yearqtr(
      ApplianceShipments$Quarter,
      format = "Q%q-%Y"
      ),
    # "frac=1" sets date to last day of quarter
    # frac = 1
    )


# Create Time Series Object
MyTS <- ts(ApplianceShipments$Shipments,
           start = c(1985, 1),
           frequency = 4)


# Plot the Time Series
autoplot(MyTS) +
  labs(title = "Appliance Shipments Over Time",
       x = "Quarter",
       y = "Shipments") +
  theme_minimal()
```



Question CH2-Q3 Answer:

Answer: Looking at the chart above that we can see traces of all four components present in the series. Some of the components are stronger than others, I would order them in terms of seasonality, trend as close contenders for strongest, followed by trend and then noise.

Chapter 3:

Performance Evaluation (Pages 67-68) #2 and 3

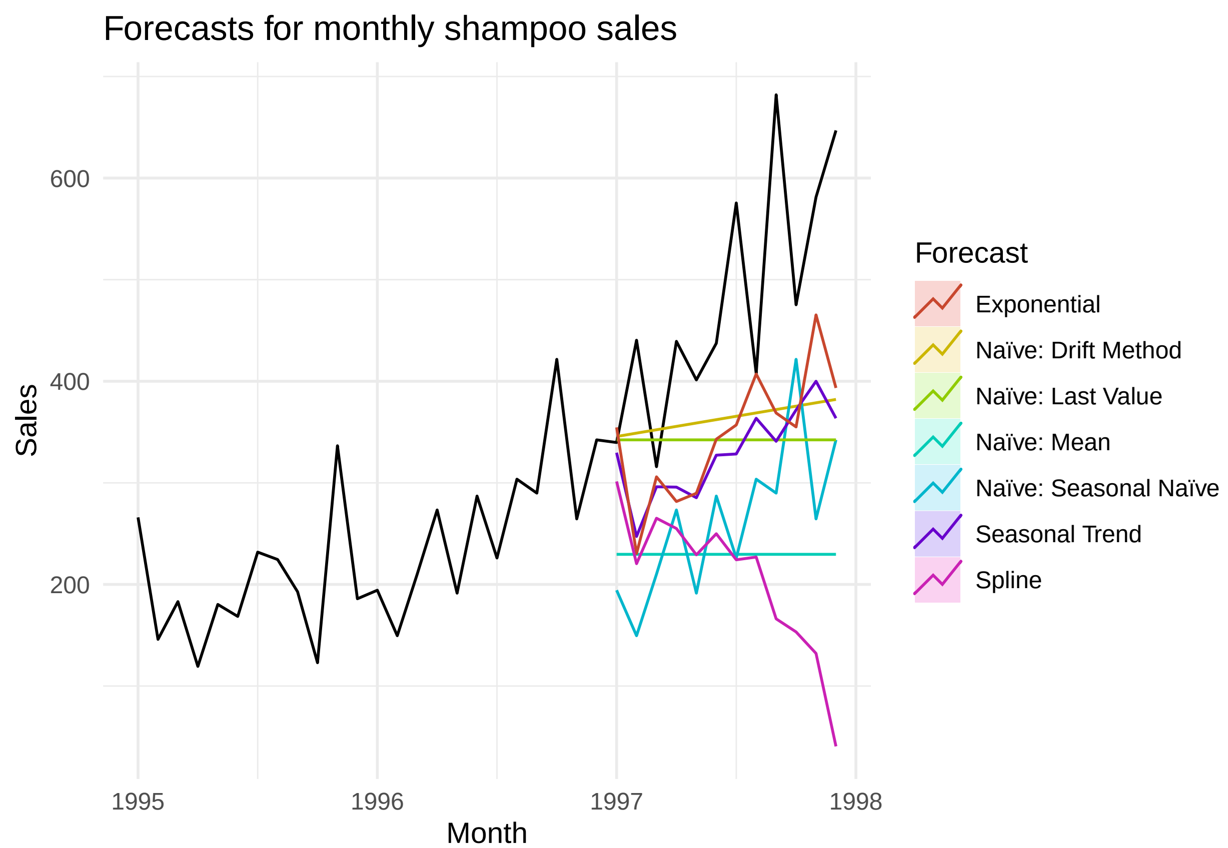
Question 2.

Forecasting Shampoo Sales: The file ShampooSales.xls contains data on the monthly sales of a certain shampoo over a three-year period. If the goal is forecasting sales in future months, which of the following steps should be taken? (choose one or more below).

- ☒ Partition the data into training and validation periods.
- ☒ Examine time plots of the series and of model forecasts only for the training period.
- ☒ Look at MAPE and RMSE values for the training period.
- ☒ Look at MAPE and RMSE values for the validation period.
- ☒ Compute naïve forecasts. Page 1 of 2

Question CH3-Q2 Answer: **Answer:** I added more models and naïve forecasts than asked for. The best model appears to be the spline or the seasonal trend model. They beat all other naïve forecasts and models in RMSE and other metrics. See my results below.

```
library(fpp2) # Plot and Forecast Data
set.seed(506)

ShampooSales <- read_excel("ShampooSales.xlsx")

ShampooSales$Month <- as.Date(ShampooSales$Month)

h=12

# Create Time Series Object
MyTS <- ts(ShampooSales$Shampoo_sales,
           start = c(1995, 1),
           frequency = 12)

# Split Data
train <- window(MyTS,
                start = c(1995, 1),
                end = c(1996, 12))


# Naïve Models
mean_fit <- meanf(train,h=h) # Naïve: Mean
naive_fit <- naive(train,h=h,level = 95) # Naïve: Last Value
drift_fit <- rwf(train,h=h,drift=TRUE) # Naïve: Drift Method
snaive_fit <- snaive(train,h=h) # Naïve: Seasonal Naïve

# Train Models
st_model <- tslm(train ~ trend + season)
st_fit <- forecast(st_model,h=h)
exp_model <- tslm(train ~ trend + season, lambda = 0)
exp_fit <- forecast(exp_model,h=h)
spline_model <- tslm(train ~ trend + I(trend^2) + I(trend^3) + season)
spine_fit <- forecast(spline_model,h=h)

# Plot Results
autoplot(window(MyTS, start=1995)) +
  autolayer(mean_fit, series="Naïve: Mean", PI=FALSE) +
  autolayer(drift_fit, series="Naïve: Drift Method", PI=FALSE) +
  autolayer(snaive_fit, series="Naïve: Seasonal Naïve", PI=FALSE) +
  autolayer(naive_fit, series="Naïve: Last Value", PI=FALSE) +
  autolayer(st_fit, series="Seasonal Trend", PI=FALSE) +
  autolayer(exp_fit, series="Exponential", PI=FALSE) +
  autolayer(spine_fit, series="Spline", PI=FALSE) +
  xlab("Month") +
  ylab("Sales") +
  ggtitle("Forecasts for monthly shampoo sales") +
  theme_minimal() +
  guides(colour=guide_legend(title="Forecast"))
```




```
window_df <- window(MyTS, start=1997)
print(strrep("#", 80))
```

```
## [1] "################################################################################"
```

```
print("Naïve: Mean")
```

```
## [1] "Naïve: Mean"
```

```
accuracy(mean_fit, window_df)
```

```
##                        ME      RMSE       MAE       MPE     MAPE      MASE
## Training set 4.149169e-15  73.46891  60.02083 -10.58964 28.85135 0.6577626
## Test set     2.489250e+02 273.05220 248.92500  49.36422 49.36422 2.7279452
##                    ACF1 Theil's U
## Training set 0.22955538        NA
## Test set     0.09989718  1.981456
```

```
print(strrep("#", 80))
```

```
## [1] "################################################################################"
```

```r
print("Naïve: Last Value")
```

```
## [1] "Naïve: Last Value"
```

```r
accuracy(naive_fit, window_df)
```

```
##                      ME      RMSE       MAE      MPE     MAPE      MASE
## Training set   3.317391  89.82947  74.07391 -6.52977 33.83955 0.8117689
## Test set     136.250000 176.51603 141.08333 24.51768 26.03809 1.5461187
##                     ACF1 Theil's U
## Training set -0.63447620        NA
## Test set      0.09989718  1.267483
```

```r
print(strrep("#", 80))
```

```
## [1] "################################################################################"
```

```r
print("Naïve: Drift Method")
```

```
## [1] "Naïve: Drift Method"
```

```r
accuracy(drift_fit, window_df)
```

```
##                        ME     RMSE       MAE       MPE     MAPE      MASE
## Training set 8.029809e-15  89.7682  74.21815 -8.142699 34.18522 0.8133496
## Test set     1.146870e+02 154.6157 121.73188 20.211607 22.41985 1.3340480
##                      ACF1 Theil's U
## Training set -0.634476204        NA
## Test set     -0.005916248  1.120773
```

```r
print(strrep("#", 80))
```

```
## [1] "################################################################################"
```

```r
print("Naïve: Seasonal Naïve")
```

```
## [1] "Naïve: Seasonal Naïve"
```

```r
accuracy(snaive_fit, window_df)
```

```
##                     ME     RMSE      MAE      MPE     MAPE     MASE       ACF1
## Training set  66.33333 121.9118  91.2500 19.00793 30.12280 1.000000 -0.3328601
## Test set     215.75833 240.4731 215.7583 43.62046 43.62046 2.364475 -0.6413139
##              Theil's U
## Training set        NA
## Test set      1.824886
```

```r
print(strrep("#", 80))
```

```
## [1] "################################################################################"
```

```r
print("Seasonal Trend")
```

```
## [1] "Seasonal Trend"
```

```r
accuracy(st_fit, window_df)
```

```
##                        ME      RMSE     MAE       MPE     MAPE      MASE
## Training set 1.361743e-14  51.14287  42.150 -5.055693 20.29458 0.4619178
## Test set     1.494250e+02 179.63441 149.425 28.364315 28.36431 1.6375342
##                    ACF1 Theil's U
## Training set -0.2833485        NA
## Test set     -0.4176140  1.366998
```

```r
print(strrep("#", 80))
```

```
## [1] "################################################################################"
```

```r
print("Exponential")
```

```
## [1] "Exponential"
```

```r
accuracy(exp_fit, window_df)
```

```
##                      ME      RMSE      MAE       MPE     MAPE      MASE
## Training set   5.352542  53.46839  40.9889 -2.364811 18.06095 0.4491934
## Test set     132.561486 165.50740 135.0412 25.016704 25.74668 1.4799038
##                    ACF1 Theil's U
## Training set -0.3454801        NA
## Test set     -0.5623278  1.288377
```

```r
print(strrep("#", 80))
```

```
## [1] "################################################################################"
```

```r
print("Spline")
```

```
## [1] "Spline"
```

```r
accuracy(spine_fit, window_df)
```

```
##                         ME      RMSE       MAE       MPE     MAPE      MASE
## Training set -5.923358e-16  45.42943  33.64367 -4.119298 15.69641 0.3686978
## Test set      2.731904e+02 322.28346 273.19043 52.063267 52.06327 2.9938678
##                    ACF1 Theil's U
## Training set -0.5415807        NA
## Test set      0.2978027  2.246434
```

```r
# This is here for my own edification.
e <- tsCV(train, rwf, drift=TRUE, h=1)
sqrt(mean(e^2, na.rm=TRUE))
```

```
## [1] 98.79881
```

```r
sqrt(mean(residuals(rwf(train, drift=TRUE))^2, na.rm=TRUE))
```

```
## [1] 89.7682
```

Question 3.

Performance on Training and Validation Data: Two different models were fit to the same time series. The first 100 time periods were used for the training period and the last 12 periods were treated as a validation period. Assume that both models make sense practically and fit the data reasonably well. Below are the RMSE values for each of the models:

Model	Training Period	Validation Period
Model A	543	690
Model B	669	675

a) Which model appears more useful for retrospectively describing the different components of this time series? Why?

b) Which model appears to be more useful for forecasting purposes? Why?

Question CH3-Q3 Answer: Answer:

a) Model A appears more useful for retrospectively describing the different components of this time series because the root mean square error is smaller than that of model B suggesting a model fit is tighter and more reflective of what transpired during that time frame.

b) Model B appears more useful for forecasting purposes because the root mean square error is smaller than that of model A. This suggests that model B generalizes better and is less over fit.