

# DSCI5340\_HW2\_SP24\_Group1

Divya Laxmi Yedanpuri Imran Nawaz Shareef Kethana Elavarthi Puja Madhav

2024-02-16

```
library(fpp3)
```

```
## -- Attaching packages ----- fpp3 0.5 --
```

```
## v tibble      3.2.1      v tsibble      1.1.3
## v dplyr       1.1.4      v tsibbledata 0.4.1
## v tidyr       1.3.1      v feasts      0.3.1
## v lubridate   1.9.3      v fable       0.3.3
## v ggplot2     3.4.4      v fabletools  0.3.4
```

```
## -- 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()
```

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method      from
##   as.zoo.data.frame zoo
```

```
search()
```

```
## [1] ".GlobalEnv"          "package:forecast"    "package:fable"
## [4] "package:feasts"      "package:fabletools"  "package:tsibbledata"
## [7] "package:tsibble"     "package:ggplot2"     "package:lubridate"
## [10] "package:tidyr"       "package:dplyr"       "package:tibble"
## [13] "package:fpp3"        "package:stats"       "package:graphics"
## [16] "package:grDevices"   "package:utils"       "package:datasets"
## [19] "package:methods"     "Autoloads"           "package:base"
```

```
us_gasoline
```

```
## # A tsibble: 1,355 x 2 [1W]
##   Week Barrels
```

```
##      <week>    <dbl>
## 1 1991 W06      6.62
## 2 1991 W07      6.43
## 3 1991 W08      6.58
## 4 1991 W09      7.22
## 5 1991 W10      6.88
## 6 1991 W11      6.95
## 7 1991 W12      7.33
## 8 1991 W13      6.78
## 9 1991 W14      7.50
## 10 1991 W15     6.92
## # i 1,345 more rows
```

Description: First, the above code chunks loads a package called “fpp3” and the “forecast” and a data set called “us\_gasoline” This data set contains information of Weekly data beginning Week 6, 1991, ending Week 3, 2017, with 1355 rows and two columns: the ‘Week’ and the supply of us\_gas ‘Barrels’.

## Subsetting the data until the end of 2004:

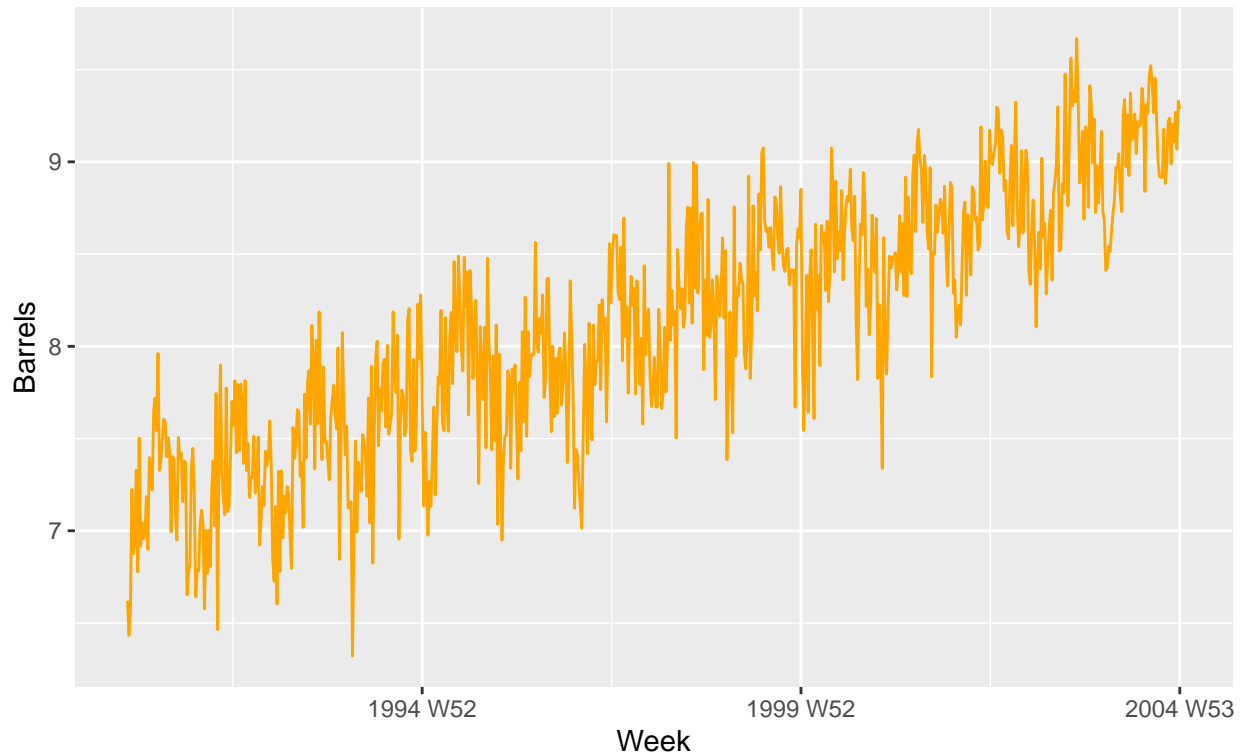
```
# Setting the training data from 1991 Week 6 to 2004 Week 53
us_gas_2004 <- us_gasoline %>%
  filter_index("1991 Week 6" ~ "2004 Week 53") %>%
  select(Week, Barrels)
us_gas_2004
```

```
## # A tibble: 726 x 2 [1W]
##       Week Barrels
##      <week>    <dbl>
## 1 1991 W06      6.62
## 2 1991 W07      6.43
## 3 1991 W08      6.58
## 4 1991 W09      7.22
## 5 1991 W10      6.88
## 6 1991 W11      6.95
## 7 1991 W12      7.33
## 8 1991 W13      6.78
## 9 1991 W14      7.50
## 10 1991 W15     6.92
## # i 716 more rows
```

```
autoplot(us_gas_2004, color = "orange" ) +
  labs(y = "Barrels",
       x = "Week",
       title = "US Gas Supply: 1991 W06 - 2004 Week53\nObserved Data")
```

```
## Plot variable not specified, automatically selected ‘.vars = Barrels’
```

## US Gas Supply: 1991 W06 – 2004 Week53 Observed Data



Description: The above code chunks subset the `us_gasoline` data set using the `filter_index()` function from Week 06, 1991 to Week 53, 2004. Next, the code uses the “`autoplot()`” function to create a graph that visualizes the data from the subsetting “`us_gas_2004`” data set. This graph helps to understand how gas supply have changed over time.

Q1. Fit a harmonic regression with trend to the above data. Experiment with changing the number Fourier terms up to 10. Select the appropriate number of Fourier terms to include by minimizing the AICc.

```
fourier_us_gas_2004_k2 <- us_gas_2004 |>
  model(TSLM(Barrels ~ trend() + fourier(K = 2)))
report(fourier_us_gas_2004_k2)
```

```
## Series: Barrels
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.9375162 -0.1897569 -0.0006692  0.2058275  1.0016928
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.094e+00  2.121e-02  334.493  < 2e-16 ***
## trend()         2.802e-03  5.057e-05  55.420  < 2e-16 ***
## fourier(K = 2)C1_52 -1.237e-01  1.497e-02  -8.265  6.71e-16 ***
## fourier(K = 2)S1_52 -2.383e-01  1.497e-02 -15.917  < 2e-16 ***
## fourier(K = 2)C2_52  4.493e-02  1.495e-02   3.006  0.00274 **
## fourier(K = 2)S2_52  1.054e-02  1.498e-02   0.704  0.48193
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.285 on 720 degrees of freedom
## Multiple R-squared:  0.8271, Adjusted R-squared:  0.8259
## F-statistic: 688.8 on 5 and 720 DF, p-value: < 2.22e-16
```

```
glance(fourier_us_gas_2004_k2) %>%
  select(adj_r_squared, CV, AIC, AICc, BIC)
```

```
## # A tibble: 1 x 5
##   adj_r_squared    CV    AIC    AICc    BIC
##   <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1      0.826 0.0819 -1815. -1814. -1783.
```

```
fourier_us_gas_2004_k3 <- us_gas_2004 %>%
  model(TSLM(Barrels ~ trend() + fourier(K = 3)))
report(fourier_us_gas_2004_k3)
```

```
## Series: Barrels
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.844141 -0.173718  0.003836  0.195973  0.908907
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.094e+00  2.063e-02 343.883 < 2e-16 ***
## trend()         2.800e-03  4.919e-05  56.921 < 2e-16 ***
## fourier(K = 3)C1_52 -1.240e-01  1.456e-02  -8.515 < 2e-16 ***
## fourier(K = 3)S1_52 -2.387e-01  1.456e-02 -16.392 < 2e-16 ***
## fourier(K = 3)C2_52  4.507e-02  1.454e-02   3.100  0.00201 **
## fourier(K = 3)S2_52  1.008e-02  1.457e-02   0.692  0.48904
## fourier(K = 3)C3_52  9.554e-02  1.456e-02   6.560 1.03e-10 ***
## fourier(K = 3)S3_52  3.709e-04  1.454e-02   0.026  0.97965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2772 on 718 degrees of freedom
## Multiple R-squared:  0.8369, Adjusted R-squared:  0.8353
## F-statistic: 526.2 on 7 and 718 DF, p-value: < 2.22e-16
```

```
glance(fourier_us_gas_2004_k3) %>%
  select(adj_r_squared, CV, AIC, AICc, BIC)
```

```
## # A tibble: 1 x 5
##   adj_r_squared    CV    AIC    AICc    BIC
##   <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1      0.835 0.0777 -1853. -1853. -1812.
```

```
fourier_us_gas_2004_k7 <- us_gas_2004 %>%
  model(TSLM(Barrels ~ trend() + fourier(K = 7)))
report(fourier_us_gas_2004_k7)
```

```
## Series: Barrels
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.8709966 -0.1723461 -0.0004569  0.1780736  0.9484129
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.095e+00  2.003e-02  354.233 < 2e-16 ***
## trend()         2.798e-03  4.776e-05  58.591 < 2e-16 ***
## fourier(K = 7)C1_52 -1.244e-01  1.413e-02  -8.804 < 2e-16 ***
## fourier(K = 7)S1_52 -2.394e-01  1.414e-02 -16.932 < 2e-16 ***
## fourier(K = 7)C2_52  4.528e-02  1.411e-02   3.208  0.0014 **
## fourier(K = 7)S2_52  9.330e-03  1.414e-02   0.660  0.5096
## fourier(K = 7)C3_52  9.626e-02  1.414e-02   6.809 2.10e-11 ***
## fourier(K = 7)S3_52  7.197e-05  1.411e-02   0.005  0.9959
## fourier(K = 7)C4_52  2.892e-02  1.413e-02   2.046  0.0411 *
## fourier(K = 7)S4_52  2.881e-02  1.412e-02   2.041  0.0416 *
## fourier(K = 7)C5_52 -3.355e-02  1.411e-02  -2.378  0.0177 *
## fourier(K = 7)S5_52  3.165e-02  1.414e-02   2.238  0.0255 *
## fourier(K = 7)C6_52 -6.565e-02  1.412e-02  -4.648 3.99e-06 ***
## fourier(K = 7)S6_52  2.803e-02  1.413e-02   1.984  0.0476 *
## fourier(K = 7)C7_52 -2.228e-02  1.414e-02  -1.576  0.1155
## fourier(K = 7)S7_52  3.274e-02  1.411e-02   2.320  0.0206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2691 on 710 degrees of freedom
## Multiple R-squared:  0.848,    Adjusted R-squared:  0.8448
## F-statistic: 264.1 on 15 and 710 DF, p-value: < 2.22e-16
```

```
glance(fourier_us_gas_2004_k7) %>%
  select(adj_r_squared, CV, AIC, AICc, BIC)
```

```
## # A tibble: 1 x 5
##   adj_r_squared    CV    AIC    AICc    BIC
##   <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1      0.845 0.0740 -1888. -1887. -1810.
```

```
fourier_us_gas_2004_k10 <- us_gas_2004 %>%
  model(TSLM(Barrels ~ trend() + fourier(K = 10)))
report(fourier_us_gas_2004_k10)
```

```
## Series: Barrels
## Model: TSLM
##
## Residuals:
```

```
##           Min           1Q           Median           3Q           Max
## -0.8625657 -0.1755251  0.0003646  0.1817594  0.9806206
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.094e+00  2.002e-02 354.363 < 2e-16 ***
## trend()          2.799e-03  4.774e-05  58.625 < 2e-16 ***
## fourier(K = 10)C1_52 -1.245e-01  1.413e-02  -8.814 < 2e-16 ***
## fourier(K = 10)S1_52 -2.395e-01  1.413e-02 -16.945 < 2e-16 ***
## fourier(K = 10)C2_52  4.529e-02  1.411e-02   3.210  0.00139 **
## fourier(K = 10)S2_52  9.203e-03  1.414e-02   0.651  0.51524
## fourier(K = 10)C3_52  9.636e-02  1.413e-02   6.819 1.98e-11 ***
## fourier(K = 10)S3_52 -4.035e-06  1.411e-02   0.000  0.99977
## fourier(K = 10)C4_52  2.905e-02  1.413e-02   2.056  0.04015 *
## fourier(K = 10)S4_52  2.884e-02  1.411e-02   2.044  0.04134 *
## fourier(K = 10)C5_52 -3.349e-02  1.410e-02  -2.375  0.01783 *
## fourier(K = 10)S5_52  3.176e-02  1.413e-02   2.247  0.02494 *
## fourier(K = 10)C6_52 -6.569e-02  1.412e-02  -4.653 3.91e-06 ***
## fourier(K = 10)S6_52  2.815e-02  1.412e-02   1.993  0.04660 *
## fourier(K = 10)C7_52 -2.240e-02  1.413e-02  -1.585  0.11340
## fourier(K = 10)S7_52  3.279e-02  1.411e-02   2.324  0.02038 *
## fourier(K = 10)C8_52 -1.671e-02  1.411e-02  -1.184  0.23676
## fourier(K = 10)S8_52 -1.432e-03  1.412e-02  -0.101  0.91926
## fourier(K = 10)C9_52 -1.768e-02  1.411e-02  -1.253  0.21066
## fourier(K = 10)S9_52 -6.335e-04  1.413e-02  -0.045  0.96424
## fourier(K = 10)C10_52 1.274e-02  1.412e-02   0.902  0.36751
## fourier(K = 10)S10_52 -2.368e-02  1.411e-02  -1.678  0.09387 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.269 on 704 degrees of freedom
## Multiple R-squared:  0.8494, Adjusted R-squared:  0.8449
## F-statistic: 189.1 on 21 and 704 DF, p-value: < 2.22e-16
```

```
glance(fourier_us_gas_2004_k10) %>%
  select(adj_r_squared, CV, AIC, AICc, BIC)
```

```
## # A tibble: 1 x 5
##   adj_r_squared    CV    AIC    AICc    BIC
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1      0.845 0.0746 -1883. -1881. -1777.
```

Description: For fitting a harmonic regression to the `us_gas_2004` data, we are experimenting with the fourier terms with `k` values from 1 to 10. Experimenting with all the `k` values, we present the AICc values for (`k = 2,3,7,10`). Among, `fourier_us_gas_2004_k2(k=2)`, `fourier_us_gas_2004_k3(k=3)`, `fourier_us_gas_2004_k7(k=7)`, and `fourier_us_gas_2004_k10(k=10)`, we conclude that `fourier_us_gas_2004_k7(k=7)` has low AICc, stating this could be a better model for all the values of `k`.

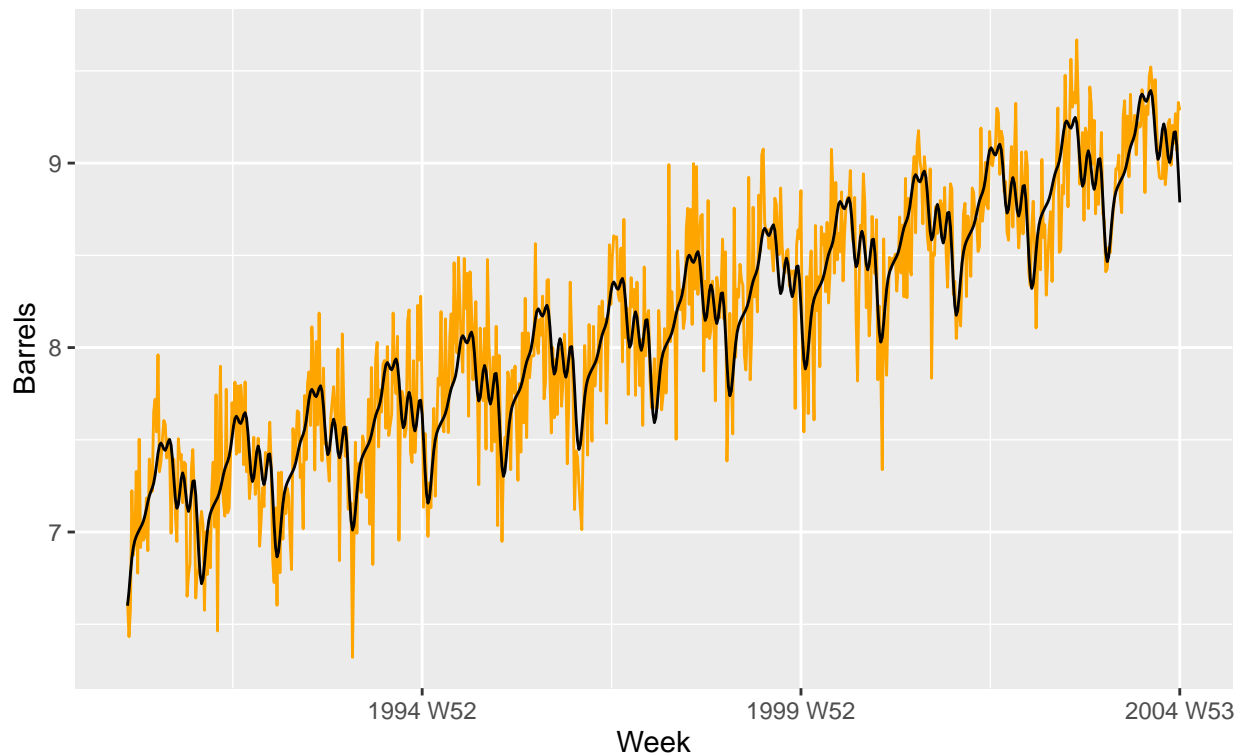
Q2. Plot the observed gasoline and fitted values. How would you characterize the fitted values?

```
us_gas_2004 %>%
  autoplot(Barrels, color='orange') +
  autolayer(fitted(fourier_us_gas_2004_k7), color='black') +
```

```
xlab("Week") + ylab("Barrels") +
ggtitle("US Gas Supply Chart with fourier_us_gas_2004_k7\nFitted Vs Observed")
```

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

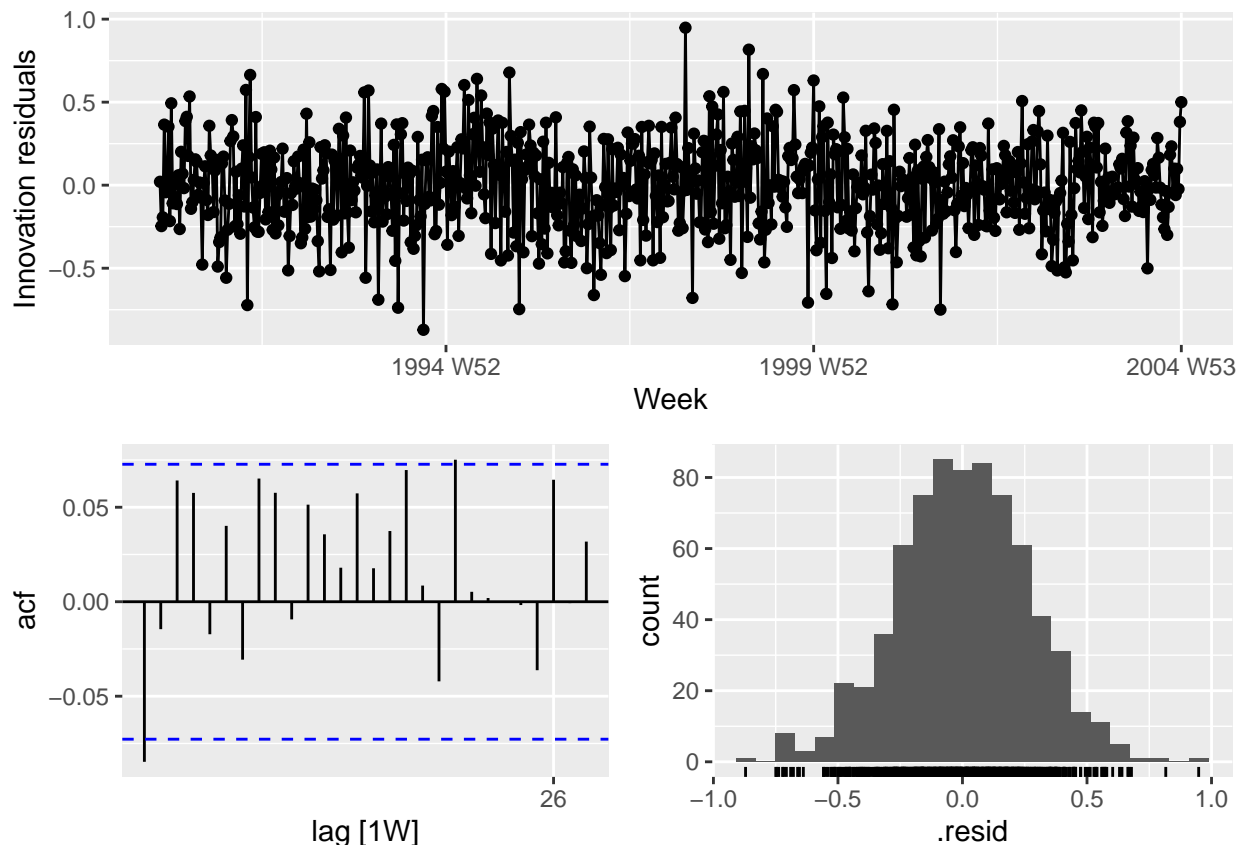
US Gas Supply Chart with fourier\_us\_gas\_2004\_k7  
Fitted Vs Observed



Description: For this question, the code is plotting the graph for “fitted value VS Observed Data” for fourier\_us\_gas\_2004\_k7, The orange line is the observed data and the black line is the fitted data, we can see that there is no much difference between the observed data pattern and the fitted data till the period of 2004.

Q3. Check the residuals of the final model using the `gg_tsresiduals()` function. Describe the patterns in various charts generated by this function.

```
fourier_us_gas_2004_k7 %>% gg_tsresiduals()
```



Description: The above code produce the residual diagnostic graphs using the `gg_tsresiduals()` function, which produces a time plot, ACF plot and histogram of the residuals. We can see that the innovation residual chart has mean close to 0 and the variance is almost constant with some outliers, but, among all the 26 vertical lines in the ACF plot, 2 of them lies outside the blue line which is more than the 5% threshold interpreting that there is auto correlation. Although, the histogram of the residual plot is normal distributed, it's not exhibiting all the properties of classifying it as a white noise, suggesting some information is still left to be picked by the model for good forecasting.

Q4. Use a Ljung-Box test to check for residual autocorrelation. What is your conclusion using the results from this test?

```
augment(fourier_us_gas_2004_k7) %>%
  features(.innov, ljung_box, lag=104, dof=5)
```

```
## # A tibble: 1 x 3
##   .model                                lb_stat lb_pvalue
##   <chr>                                <dbl>   <dbl>
## 1 TSLM(Barrels ~ trend() + fourier(K = 7)) 144.    0.00207
```

Description: We are using the Ljung-Box test to check for residual autocorrelation for the `fourier_us_gas_2004_k7` term, and got the p-value 0.002071773, which is less than 0.05, it suggests that there is evidence to reject the null hypothesis that the residuals are uncorrelated (white noise). In other words, a low p-value indicates the presence of residual autocorrelation, interpreting that the model has not picked all the information required for forecasting the supply of `us_gas_2004`, so, we need to do further analysis.

Q5. Generate forecasts for the next year of data. Plot these along with the actual data for 2005. How good are these forecasts? Explain.



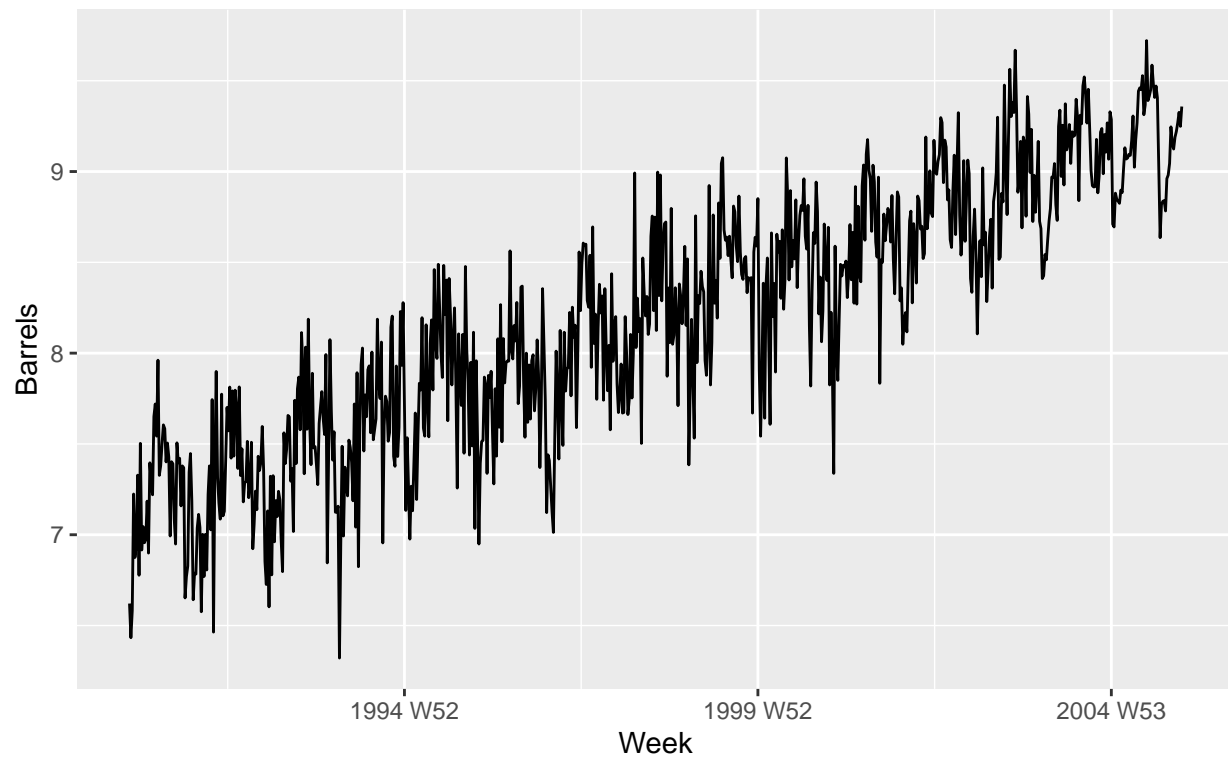
```
us_gas_2005 <- us_gasoline %>%
  filter_index("1991 W06" ~ "2005 W52") %>%
  select(c(Week, Barrels))
us_gas_2005
```

```
## # A tsibble: 778 x 2 [1W]
##       Week Barrels
##   <week>   <dbl>
## 1 1991 W06   6.62
## 2 1991 W07   6.43
## 3 1991 W08   6.58
## 4 1991 W09   7.22
## 5 1991 W10   6.88
## 6 1991 W11   6.95
## 7 1991 W12   7.33
## 8 1991 W13   6.78
## 9 1991 W14   7.50
## 10 1991 W15  6.92
## # i 768 more rows
```

```
autoplot(us_gas_2005, color = "black" ) +
  labs(y = "Barrels",
       x = "Week",
       title = "US Gas Supply: 1991 W06 - 2005 Week52\nObserved Data")
```

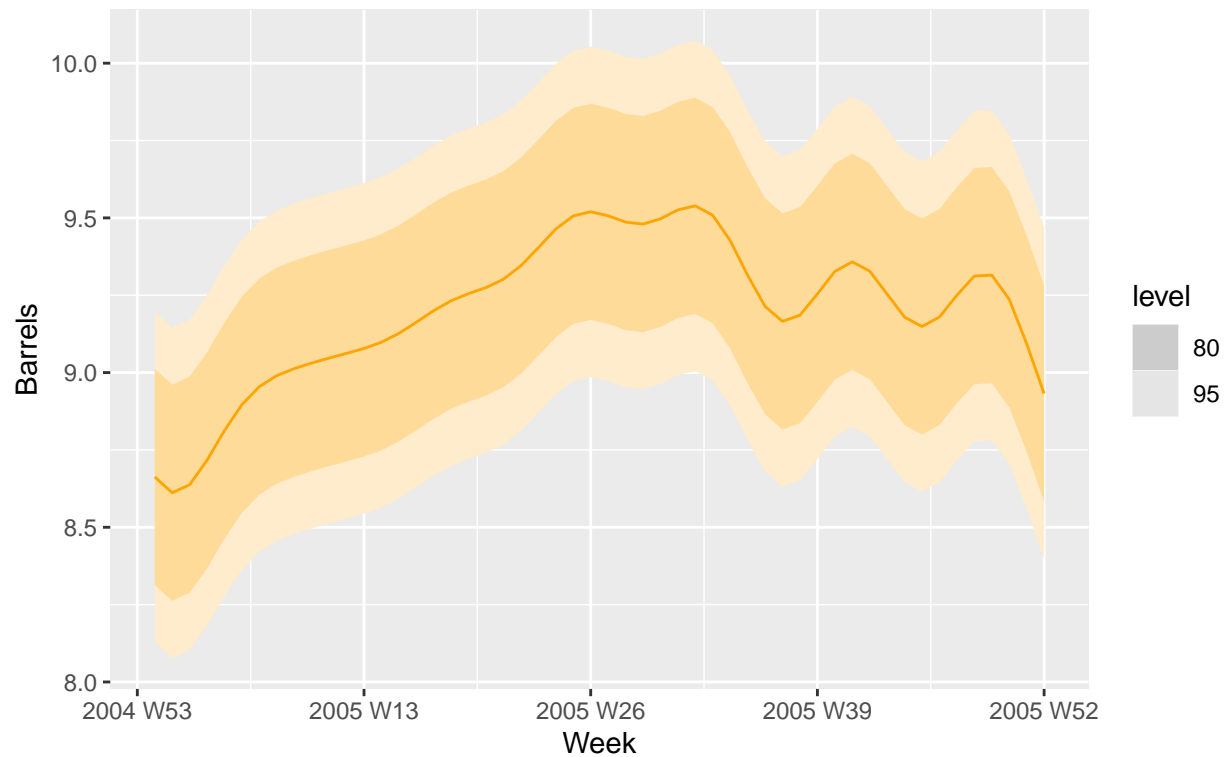
```
## Plot variable not specified, automatically selected '.vars = Barrels'
```

## US Gas Supply: 1991 W06 – 2005 Week52 Observed Data



```
# Generate forecasts for next 52 weeks
us_gas_2005_fc <- fourier_us_gas_2004_k7 %>% forecast(h = 52)
autoplot(us_gas_2005_fc, color = "orange" ) +
  labs(y = "Barrels",
       x = "Week",
       title = "US Gas Supply: 2005W01 - 2005 Week52\nForecast Graph")
```

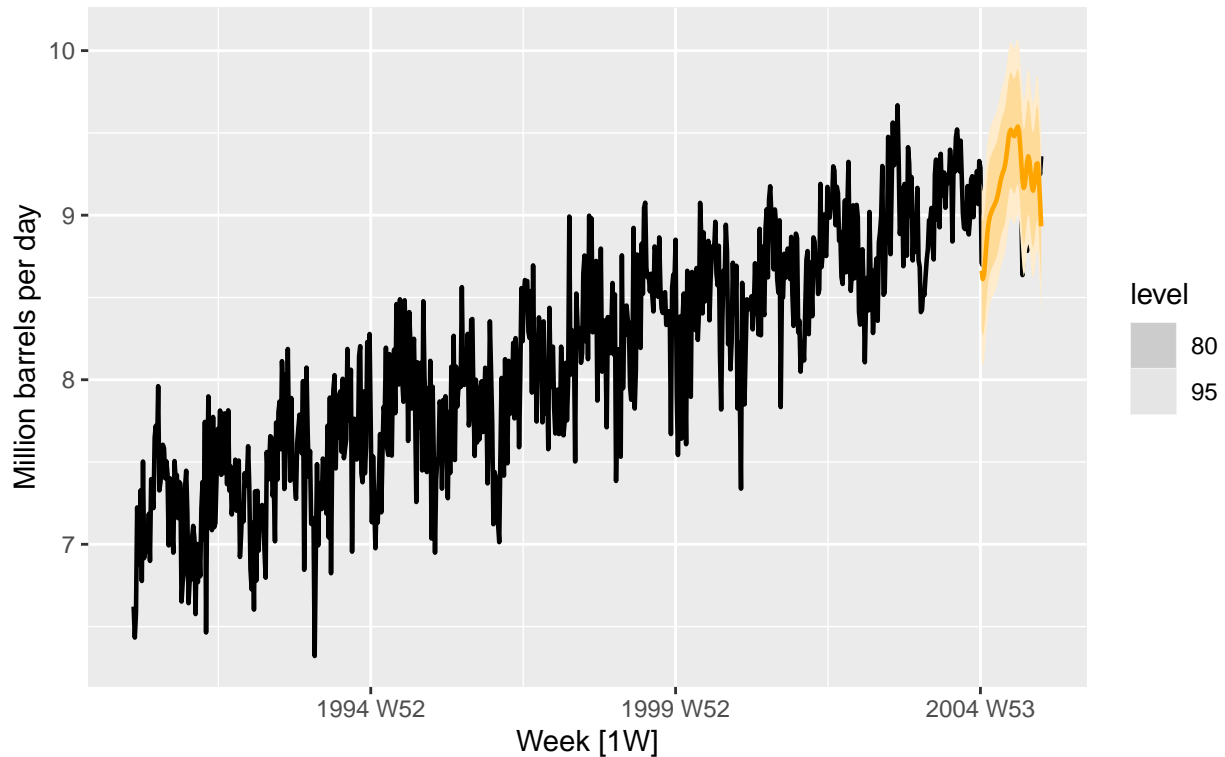
## US Gas Supply: 2005W01 – 2005 Week52 Forecast Graph



```
# Plotting the forecast for 2005 W52:
us_gas_2005 %>%
  autoplot(cex = 0.8, color = 'black') +
  labs(y = "Million barrels per day",
       title = "US Gasoline Supply: Forecast for 2005\n2005: Observed Vs Forecast") +
  autolayer(us_gas_2005_fc, cex = 0.8, color = 'orange')
```

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

## US Gasoline Supply: Forecast for 2005 2005: Observed Vs Forecast



Description: The above code chunks give the data from the year 1991 Week 1 to Week 52, 2005. We can see that after the year 2005 has been included in the data, the observations increases to 778. The 1st chart uses the `autoplot()` function to plot the graph with actual data from the year 1991 to 2005. We generate the forecast for 2005, taking the value of  $h = 52$  and forecast graph has been plotted to see the variation. We plot the graph “Observed VS Forecast” to see the deviation of the forecast line from the observed value. We can say that there is no much difference between the observed and the forecast line drawn for the year 2005 and how the forecast has captured the seasonal pattern from the observed data. Each future value is expected to lie within the 80% to 95% prediction interval.