Problem 1

Part a

Starting our analysis in the first question with usual autoplot in Fig-1. It seems like there is a logarithmic fluctuation on seasonality. Also, there is an upward trend in the data.

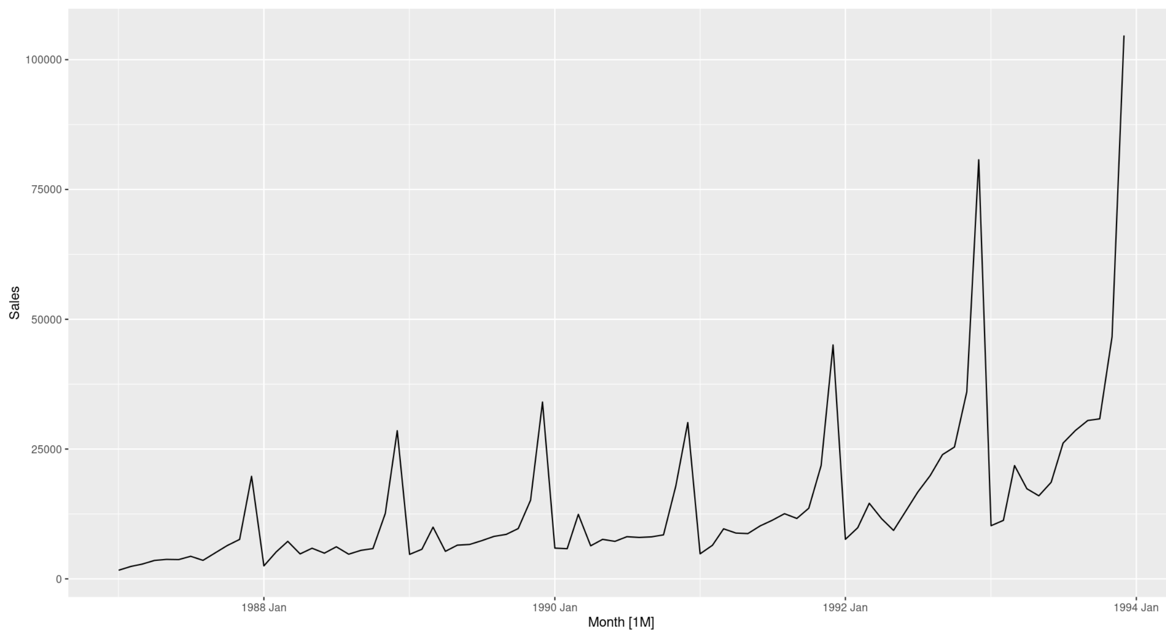

Figure 1: Monthly souvenirs sales at beach resort town in Queensland, Australia. Large influx during Christmas and local surfing festival, possible seasonality on March expect 1987

As I said in Fig-1, there is a seasonal fluctuations in the data, so we applied logarithmic to the sales to get an additive model and stabilize the variance in seasonality.

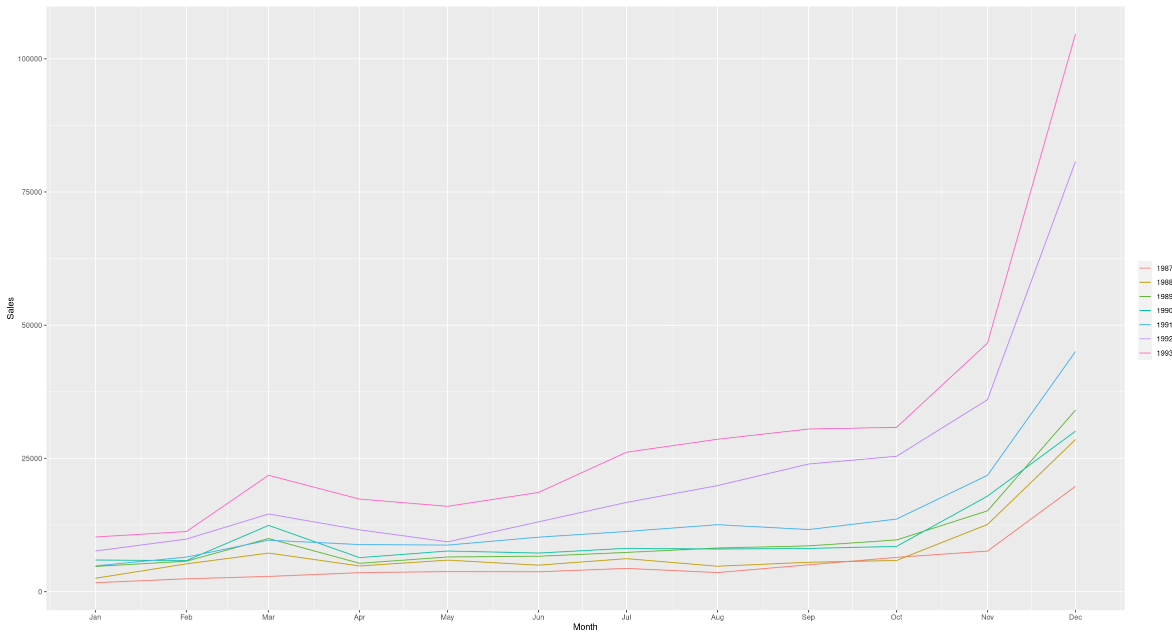

Figure 2: Seasonal plot for the same data. Seasonal patterns on March (except 1987) due to the surfing festival, and around December due to Christmas are evident.

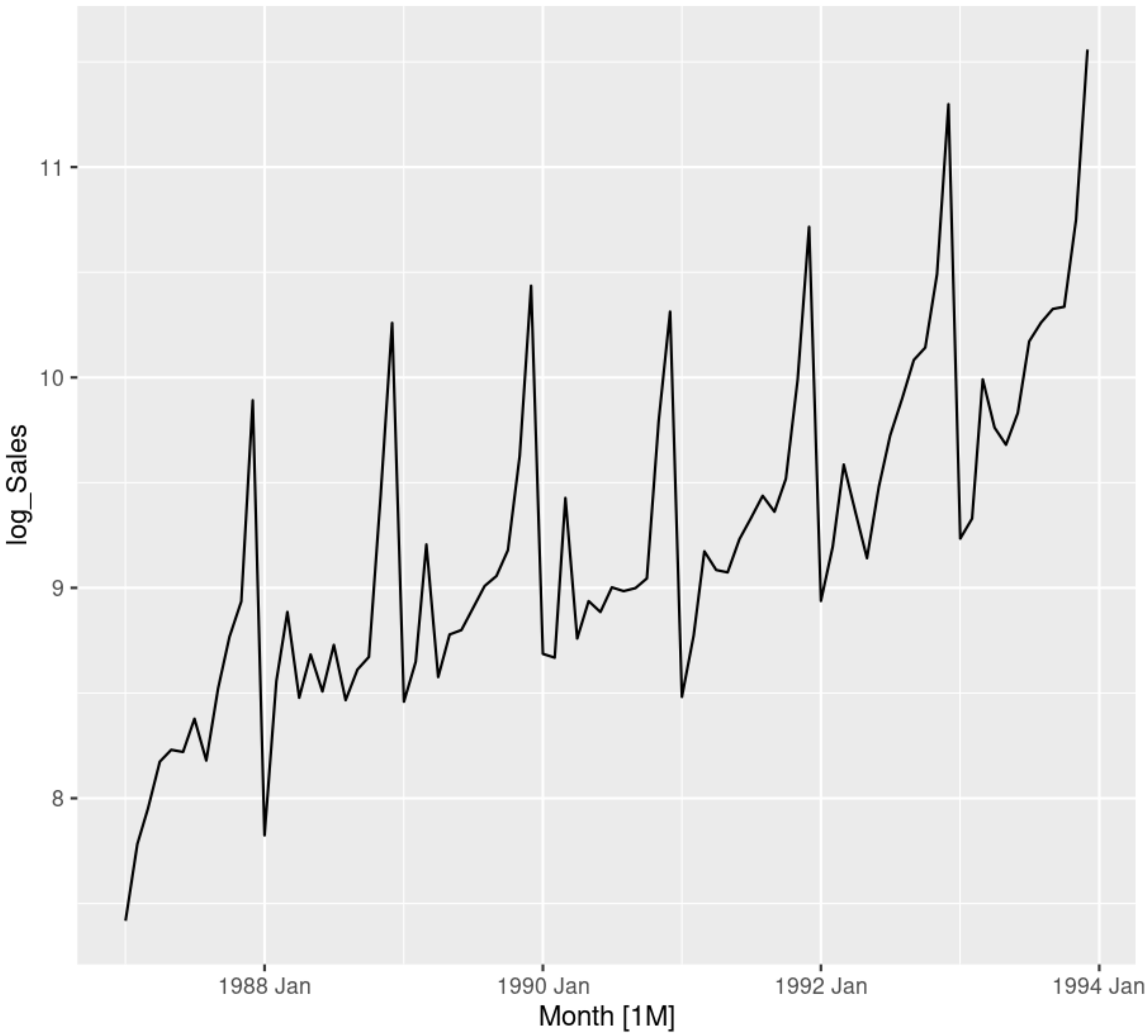

Figure 3: Sales on logarithmic scale. That reduces the effect due to seasonal fluctuations and keep the seasonal variation constant.

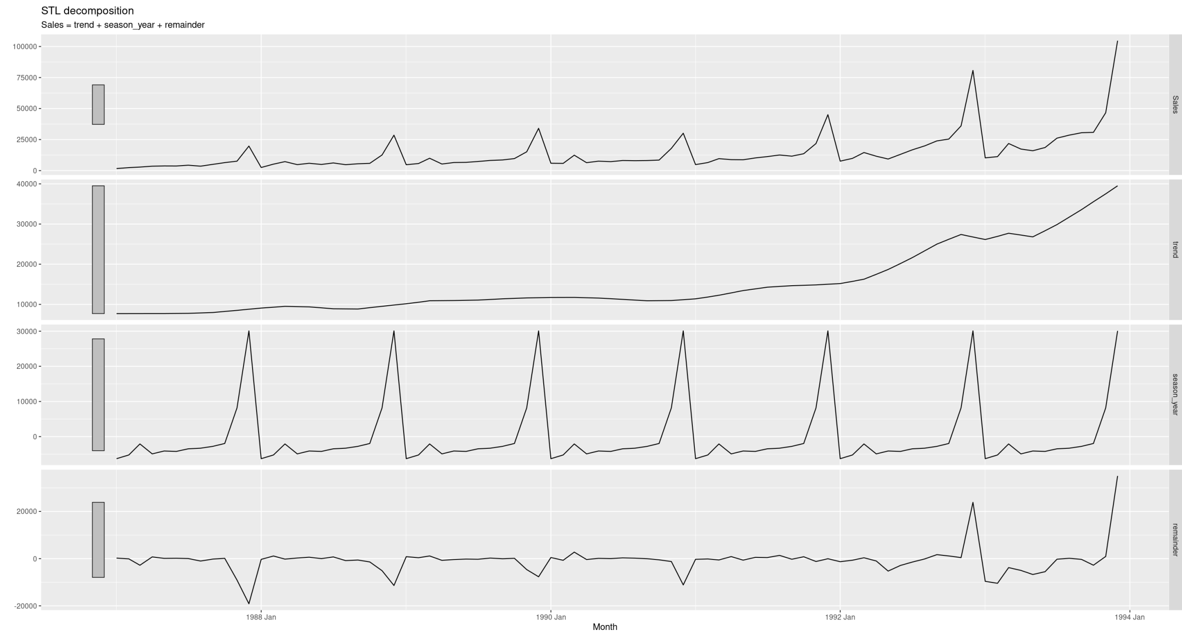

Figure 4: STL decomposition is performed on data and provided trend and stable seasonal patterns.

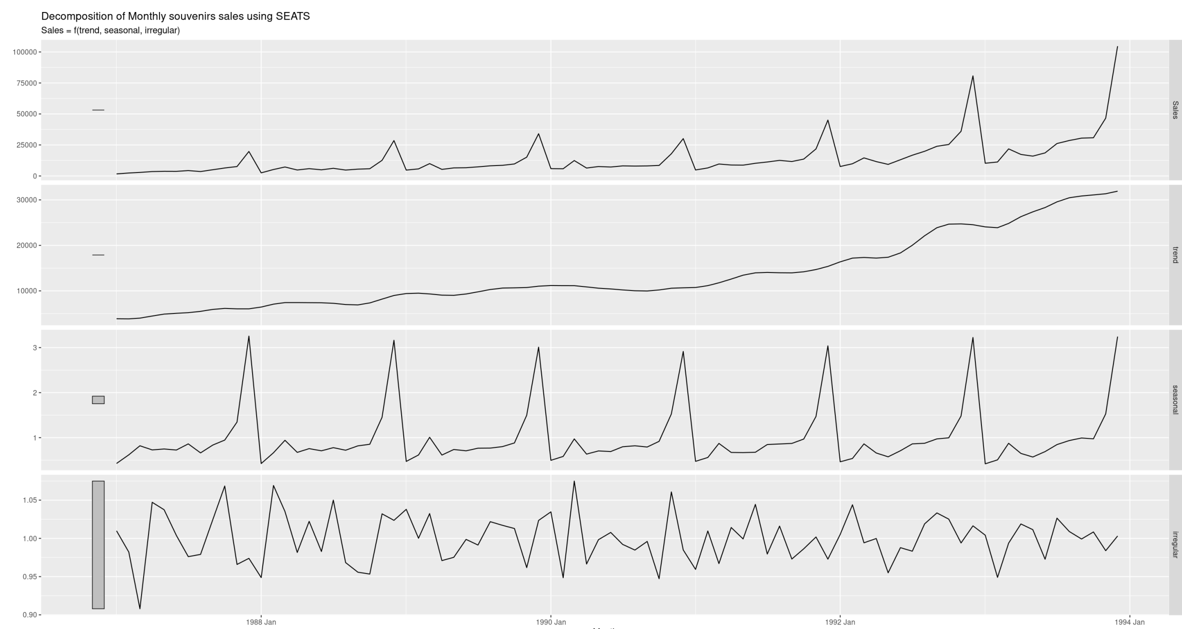

Figure 5: SEATS decomposition is performed on data set. Unlike STL method, even though seasonal pattern has less stable amplitudes, SEATS has lower interval for irregular part.

Part b

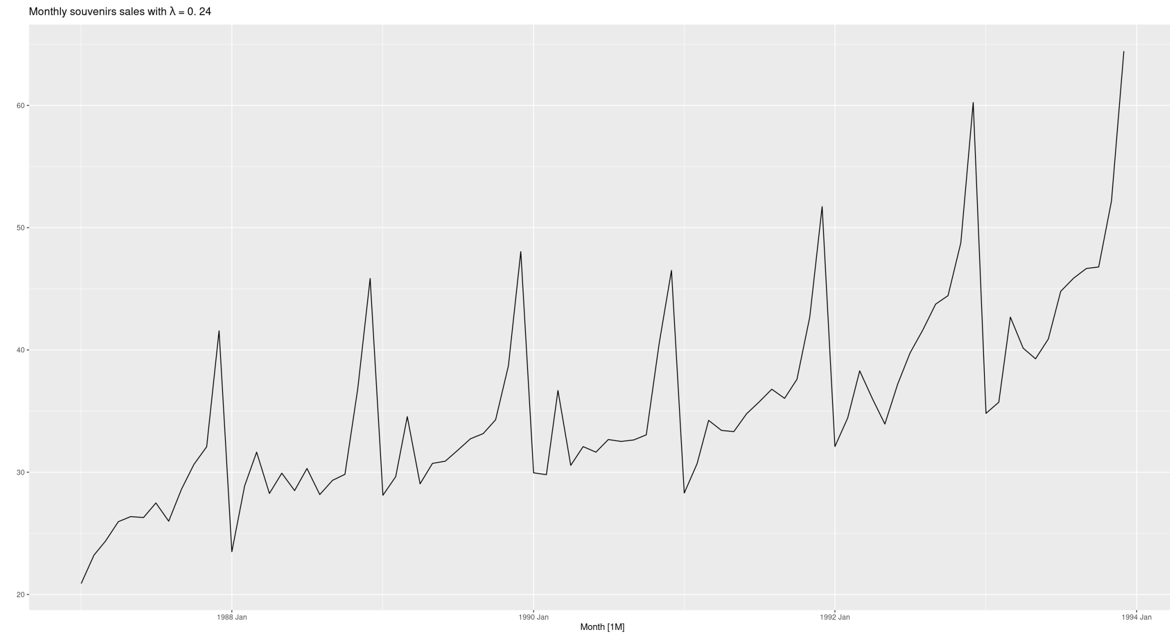

Figure 6: Guerrero method to apply Box-cox transformation ($\lambda = 0.24$) to souvenirs sales in Australia.

Part c

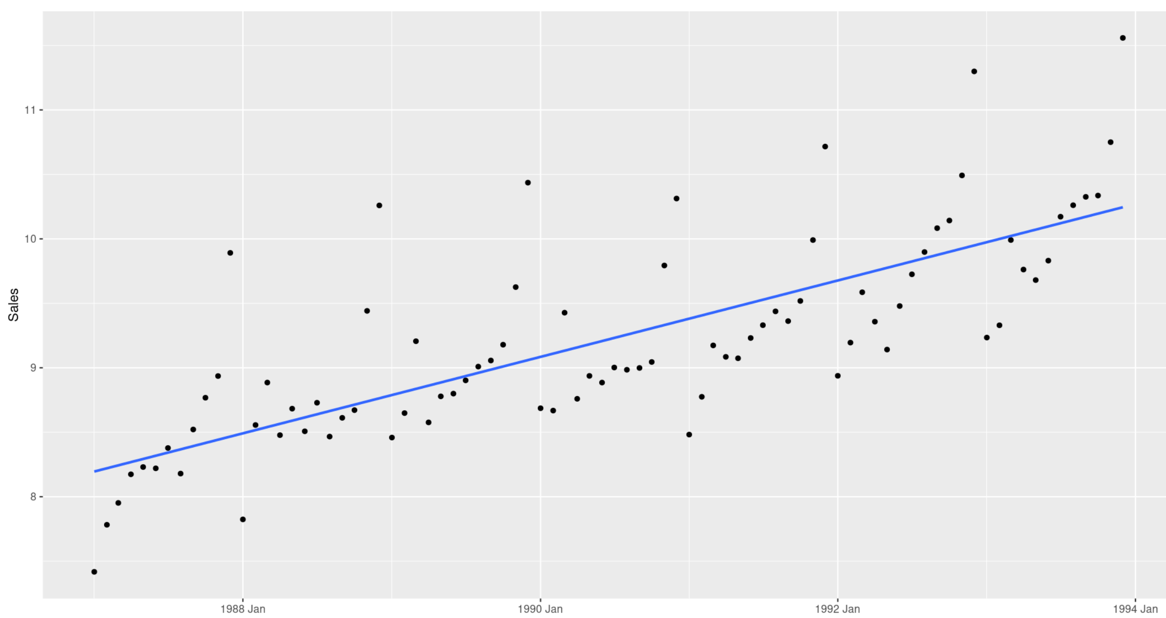

Figure 7: Regression model fit to the logarithm of sales on the data.

Part d

In part d of first question, we see residual analysis of the souvenirs sales data compare to the fitted values and time. It seems like there are still some information in residuals since there are big spikes deviated from mean in resid vs time data. In addition to this residuals show big fluctuations after 25000 in fitted value on the upper panel.

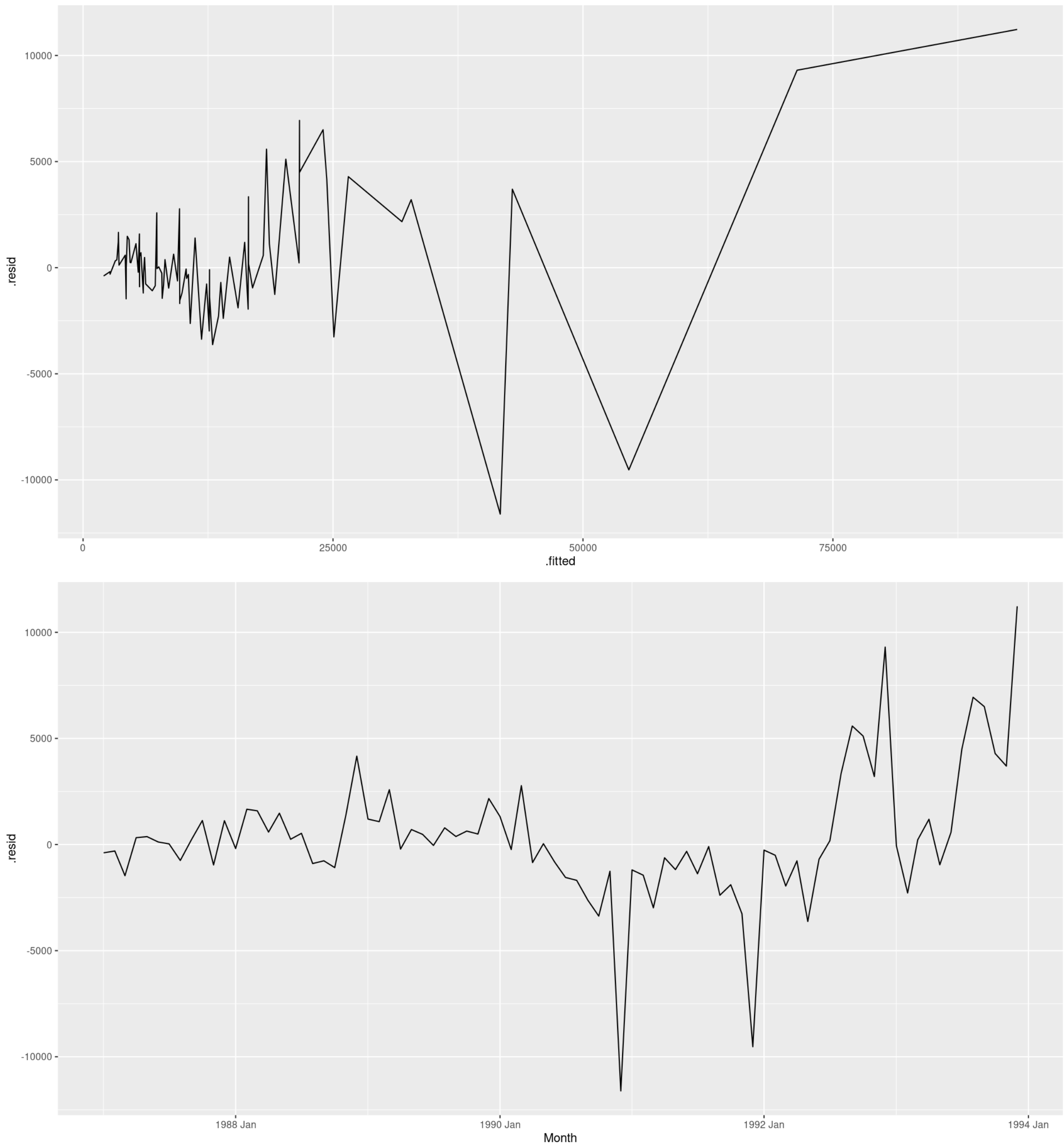

Figure 8: Time series residuals and histogram for souvenirs data.

Part e

According to the Fig-9, although error bars oscillates between -0.4 and 0.4, means of individual months do not deviate much from the mean value of the residuals.It appears as there are no problem with our model since there are not much difference from mean of the residuals for each month.

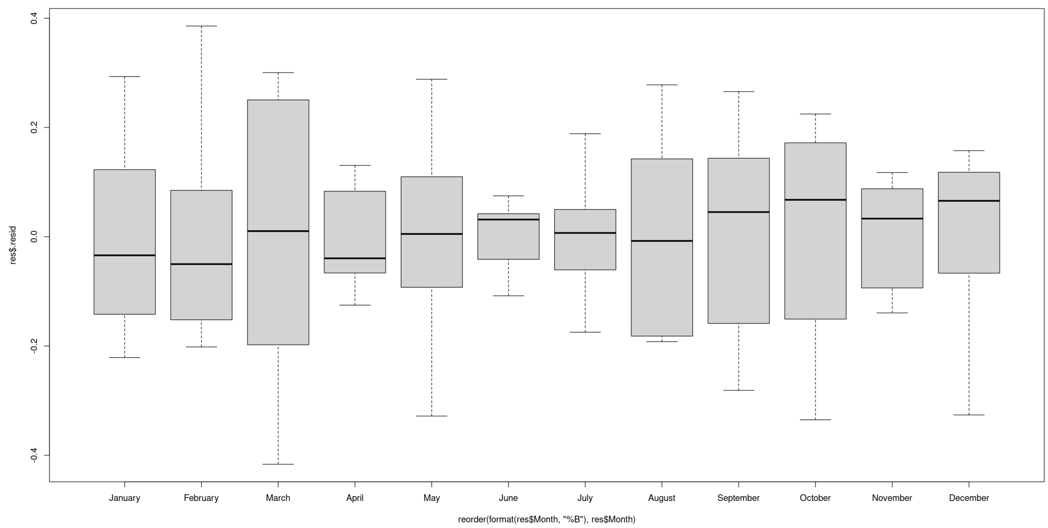

Figure 9: Box plot of residuals for each months.

Part f

In Fig-10, standard errors each years are almost the same while estimates range from 0.0224 to 1.96. P-values are generally very small, and getting smaller through the years.

```
> ##### PART F #####
> # check coeffs
> coef(fit_svnr)
# A tibble: 13 × 6
   .model                                   term            estimate std.error statistic  p.value
   <chr>                                    <chr>              <dbl>     <dbl>     <dbl>    <dbl>
 1 TSLM(log(Sales) ~ trend() + season()) (Intercept)        7.61     0.0769       98.9  8.19e-78
 2 TSLM(log(Sales) ~ trend() + season()) trend()            0.0224   0.000845     26.5  1.48e-38
 3 TSLM(log(Sales) ~ trend() + season()) season()year2      0.251    0.0993        2.53 1.37e- 2
 4 TSLM(log(Sales) ~ trend() + season()) season()year3      0.695    0.0993        7.00 1.18e- 9
 5 TSLM(log(Sales) ~ trend() + season()) season()year4      0.383    0.0994        3.85 2.52e- 4
 6 TSLM(log(Sales) ~ trend() + season()) season()year5      0.408    0.0994        4.11 1.06e- 4
 7 TSLM(log(Sales) ~ trend() + season()) season()year6      0.447    0.0994        4.50 2.63e- 5
 8 TSLM(log(Sales) ~ trend() + season()) season()year7      0.608    0.0995        6.12 4.69e- 8
 9 TSLM(log(Sales) ~ trend() + season()) season()year8      0.585    0.0995        5.88 1.21e- 7
10 TSLM(log(Sales) ~ trend() + season()) season()year9      0.666    0.0996        6.69 4.27e- 9
11 TSLM(log(Sales) ~ trend() + season()) season()year10     0.744    0.0996        7.47 1.61e-10
12 TSLM(log(Sales) ~ trend() + season()) season()year11     1.20     0.0997       12.1  7.22e-19
13 TSLM(log(Sales) ~ trend() + season()) season()year12     1.96     0.0998       19.6  2.12e-30
```

Figure 10: Coefficients for TSLM model with trend and season.

Part g

Ljung box test shows any of a group of auto-correlations of a time series are different from zero up to a specific lag. Statistics and p-values has been shown in Fig=11. P-value is extremely small, which tells us to reject the null hypothesis since there is at least one auto-correlation which is not zero in population.

```
> ##### PART G #####
> augment(fit_svnr)%>% features(.innov, ljung_box, lag = 10, dof = 0)
# A tibble: 1 × 3
  .model                                 lb_stat lb_pvalue
  <chr>                                    <dbl>     <dbl>
1 TSLM(log(Sales) ~ trend() + season())     64.7  4.60e-10
```

Figure 11: Ljung Box test for souvenirs sales data.

Part h

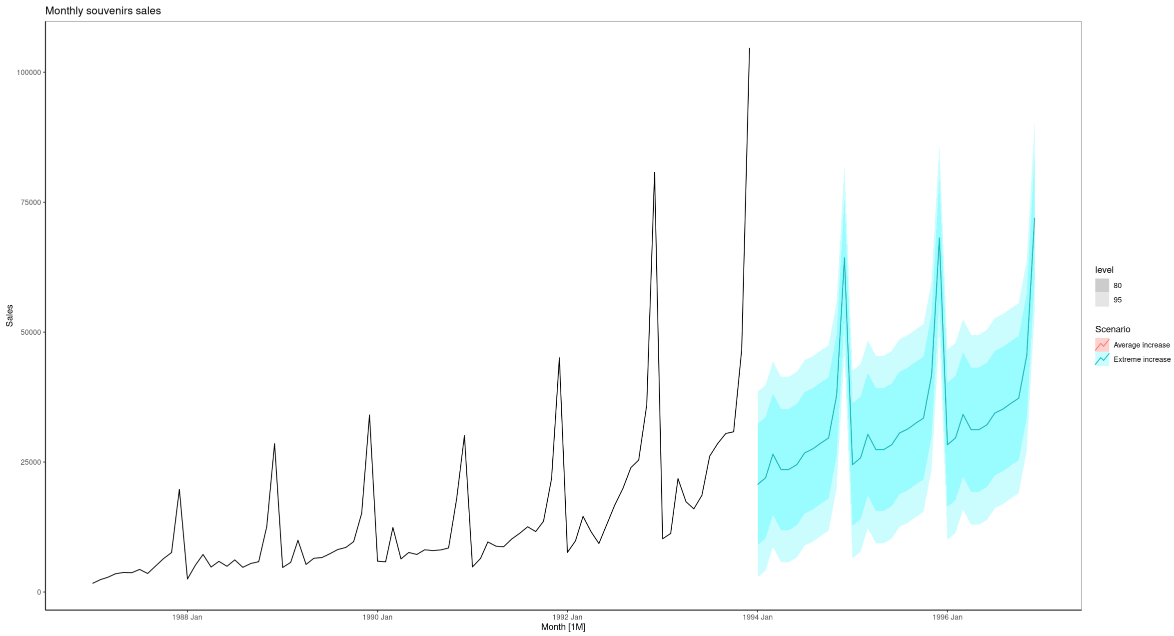

Figure 12: Predictive regression model on monthly sales for 1994, 1995, and 1996.

Part i

Last figure of this question related to interpretation of Fig-12. First thing I noticed is that our forecast does not affected by seasonal fluctuations and compare to the data, our predictions for each year show a realistic interval for the forecast following the trend.

Problem 2

Fig-13 represents that with increasing number of Fourier term, our fit show better performance on following the trend and showing detailed features in the data. Even though trend is easy to see in the data for each Fourier terms, cyclic behaviour is harder to catch and more accurate fits of data provide better resolution for cyclic behaviour.

Part a

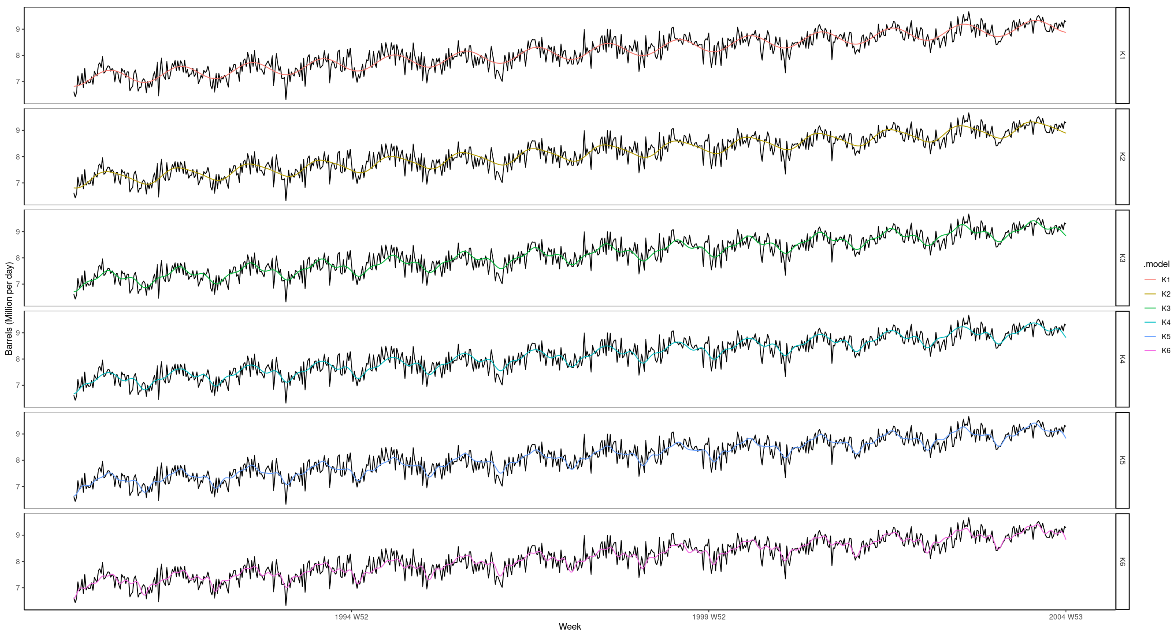

Figure 13: Harmonic regression fit with trend to the data. Fourier term changes between 1-6.

Part b

Various values have been tested for different number of Fourier terms. AICc exhibits that the higher number of Fourier terms, the better AICc value. However, we see the lowest value at K=5, not K=6. K=5 is our turning point in terms of AICc value which shows the best fit. Moreover, CV test tells us again the lowest value, K=5, has the best fit for our data. On the other hand, logarithmic likelihood displays that last two fits have the same result while others indicate K=5 is the optimum fit for the data.

```
> glance(fit) %>%
+   select(.model, sigma2, log_lik, AIC, AICc, BIC, CV)
# A tibble: 6 × 7
  .model    sigma2 log_lik    AIC   AICc    BIC      CV
  <chr>      <dbl>   <dbl>  <dbl>  <dbl>  <dbl>   <dbl>
1 K1     0.00232      292. -1086. -1085. -1070. 0.00238
2 K2     0.00213      301. -1100. -1099. -1077. 0.00220
3 K3     0.00150      334. -1161. -1160. -1132. 0.00157
4 K4     0.00130      348. -1185. -1183. -1149. 0.00138
5 K5     0.000966     376. -1236. -1234. -1195. 0.00104
6 K6     0.000969     376. -1235. -1232. -1190. 0.00105
```

Figure 14: Different test results for TSLM model.

Part c

The figures through 15-20 show residual analyses for the different Fourier terms. Innovation residuals are between $-1, +1$ and look randomly distributed along the years. Auto-correlation function provides to valuable information about lags, in which there 3-4 lags in each Fourier terms show there are still some level of valuable information in residuals. Also note that, K=6 gives the best auto-correlation result, which almost shows a pure white noise. In histograms, each plot shows us almost a Gaussian distribution, little skewed to the right or left, depending on number of Fourier terms. It is crucial to add, there are also small number of outliers in each Fourier terms.

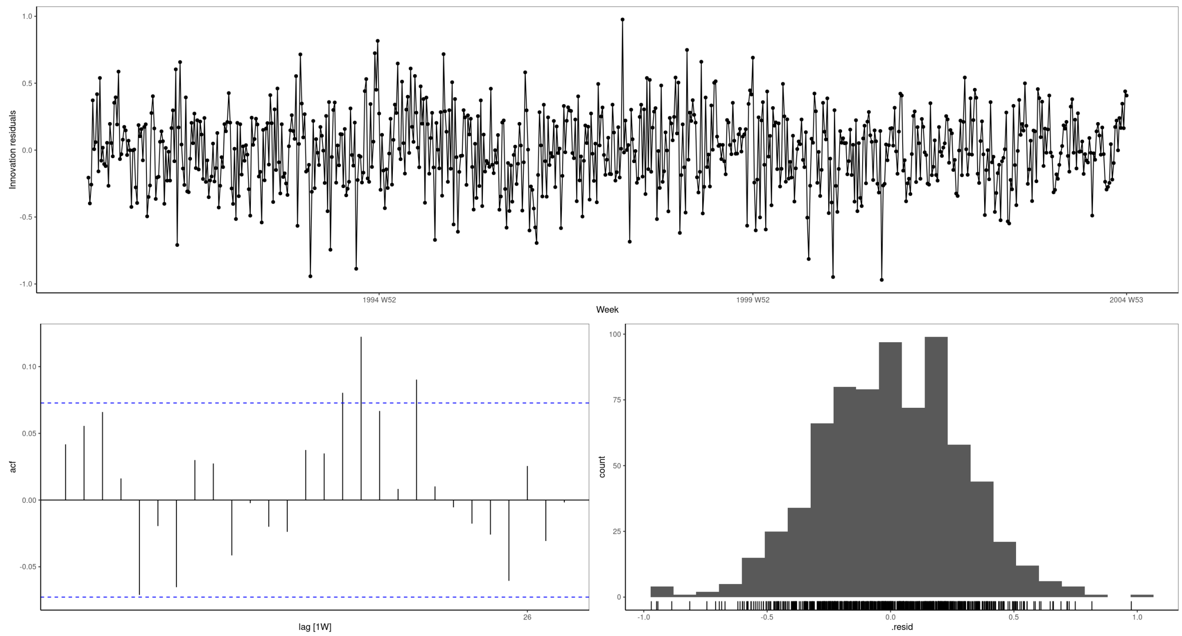

Figure 15: Auto-correlation and histogram for residuals of Fourier term equals 1.

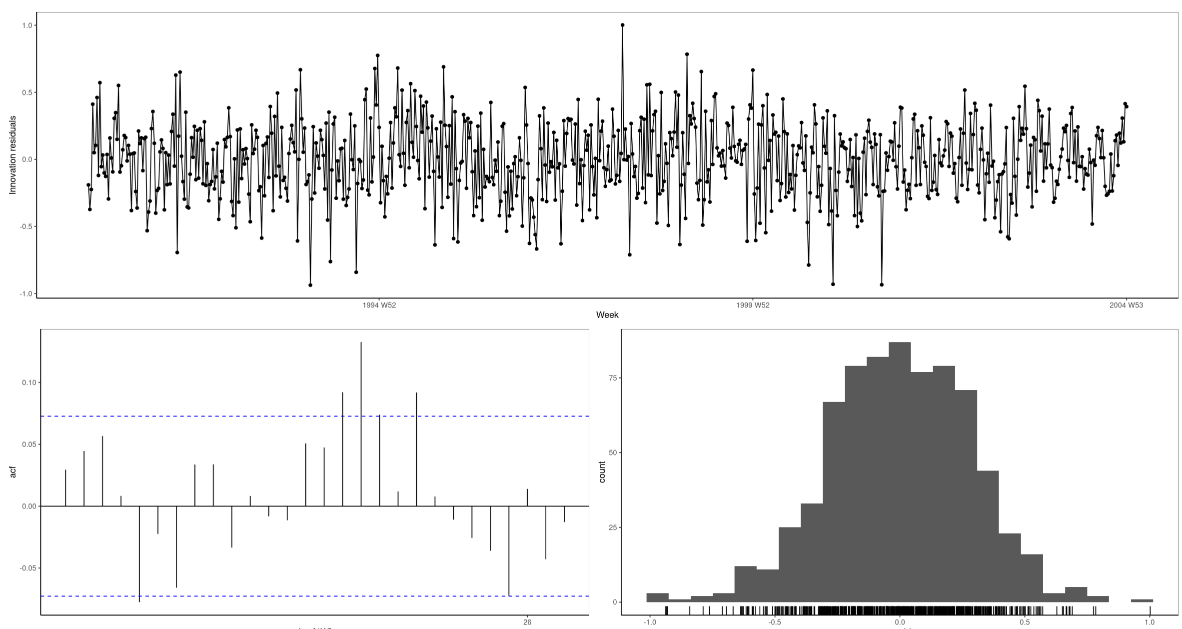

Figure 16: Auto-correlation and histogram for residuals of Fourier term equals 2.

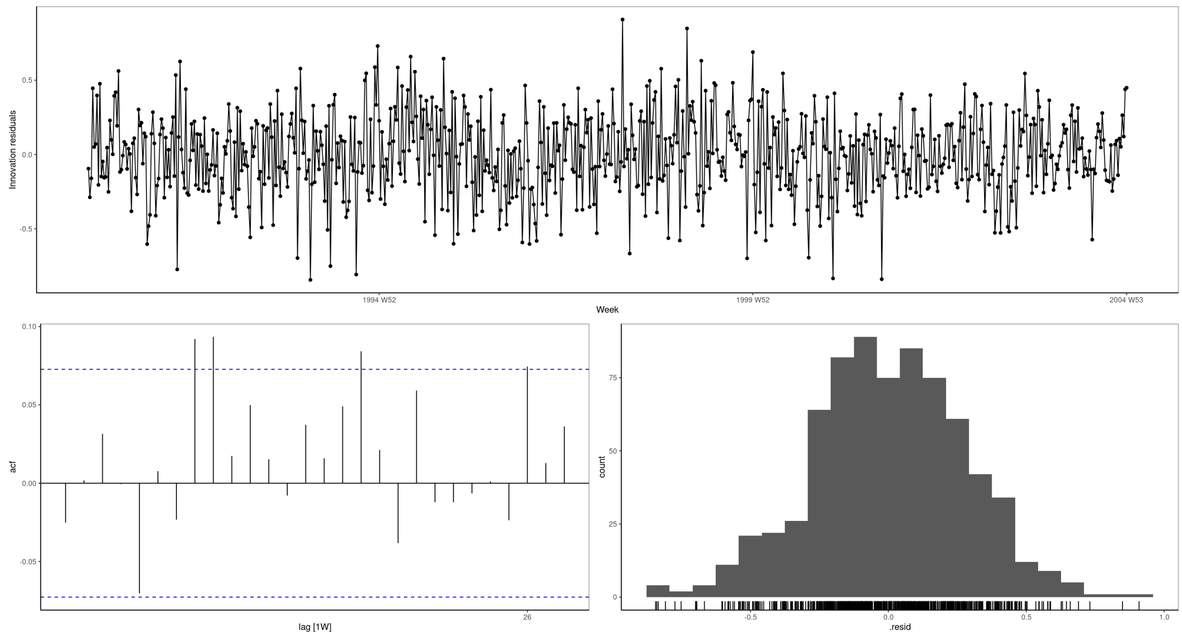

Figure 17: Auto-correlation and histogram for residuals of Fourier term equals 3.

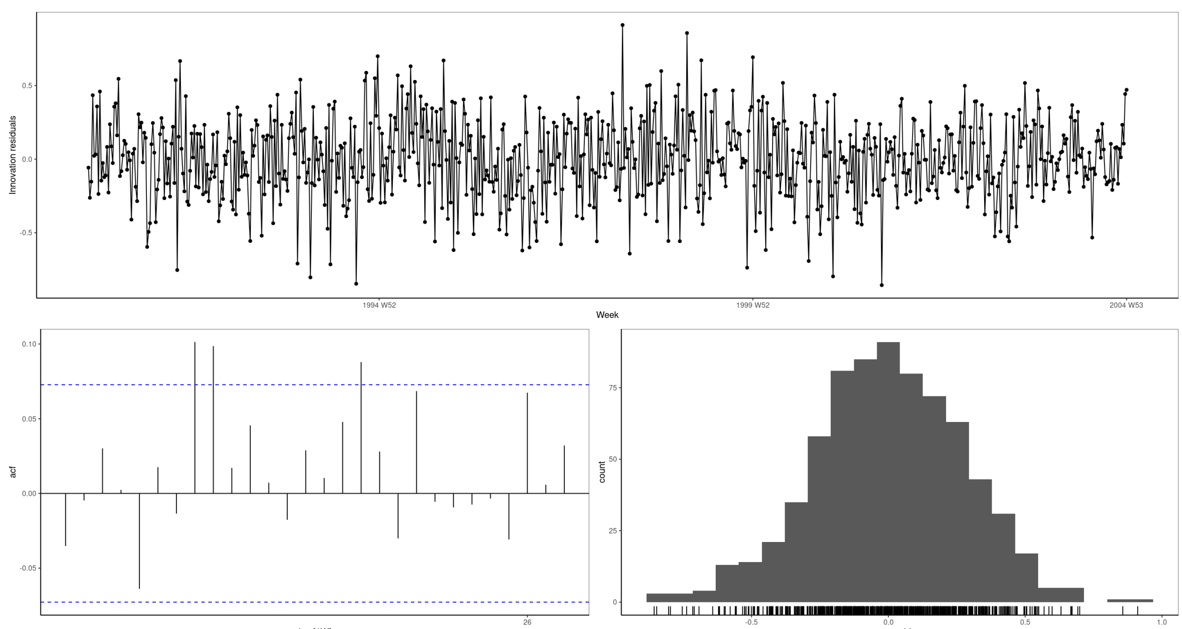

Figure 18: Auto-correlation and histogram for residuals of Fourier term equals 4.

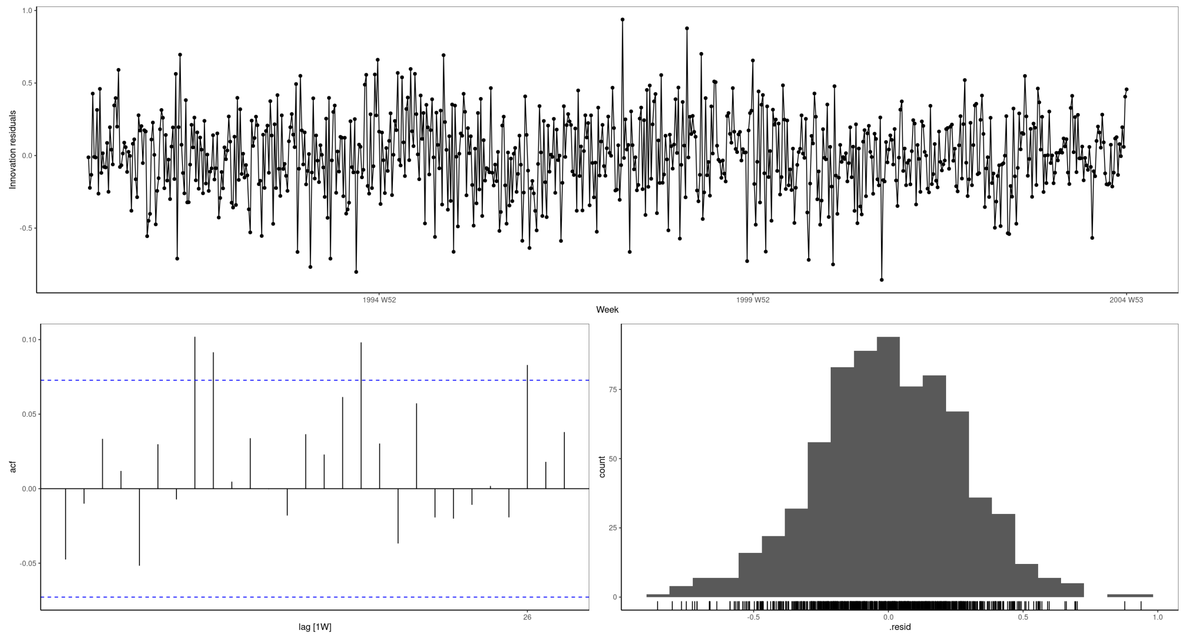

Figure 19: Auto-correlation and histogram for residuals of Fourier term equals 5.

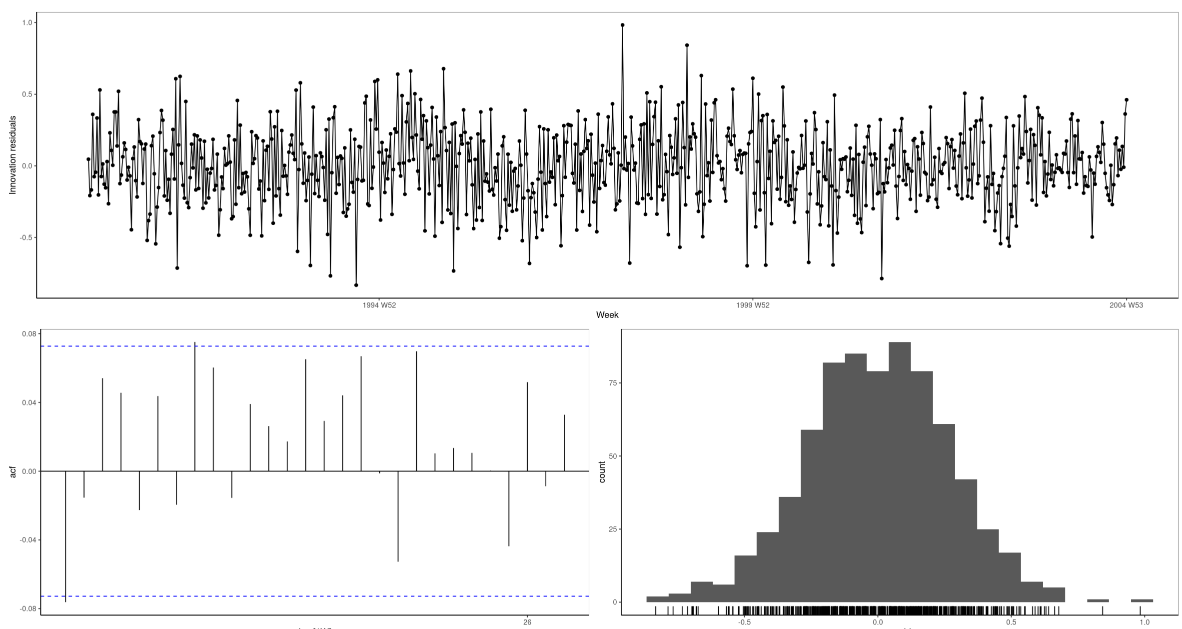

Figure 20: Auto-correlation and histogram for residuals of Fourier term equals 6.

Ljung box test for the first 10 lags on different Fourier terms show similar results, except K=2.

```
> fit_usgas %>% augment()%>% features(.innov, ljung_box, lag = 10, dof = 0)
# A tibble: 6 × 3
  .model lb_stat lb_pvalue
  <chr>    <dbl>     <dbl>
1 K1        16.5    0.0861
2 K2        15.0    0.133
3 K3        18.2    0.0522
4 K4        19.9    0.0305
5 K5        19.1    0.0387
6 K6        17.1    0.0714
```

Figure 21: Ljung-Box test for different Fourier terms.

Part d

I provide 3 different forecasts, linear, exponential and piecewise, for 2005. while linear and piecewise forecasts show exactly the same fit and prediction for 2005, exponential forecast shows slightly higher prediction. Even though exponential forecast shows better prediction on last week of the year, linear and piecewise forecasts show narrower prediction interval in 2005. As a result, we can say linear and piecewise forecasts have better result.

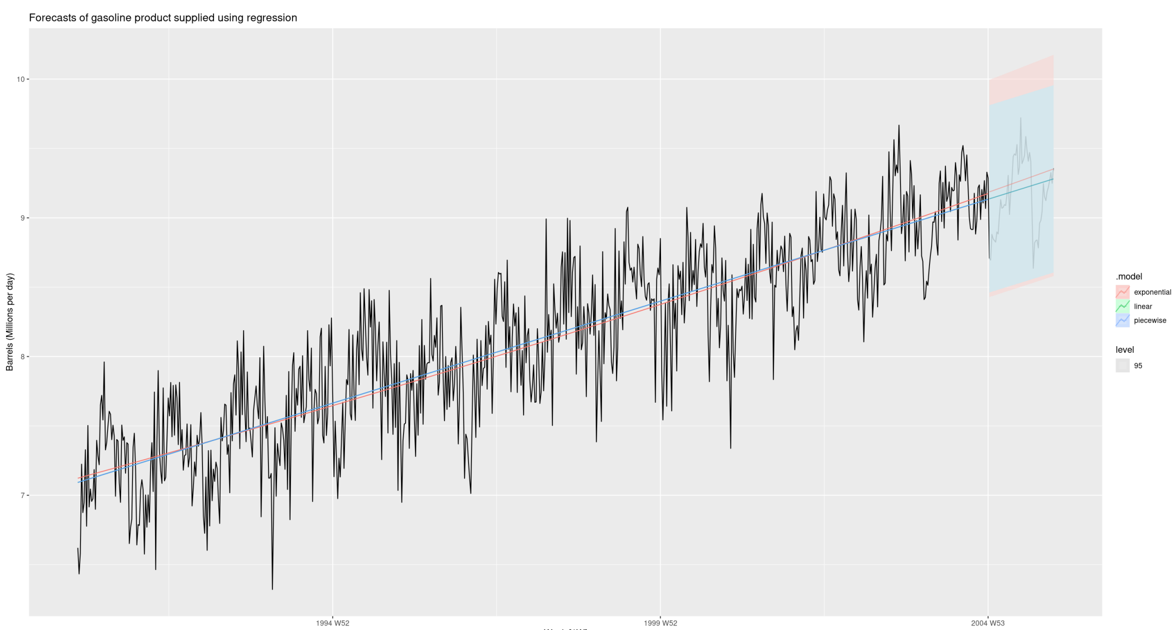

Figure 22: Exponential, linear and piecewise forecasts for supplies of US finished motor gasoline product.

On the other hand, regression forecasting, compare the previous forecasts, shows us much more better prediction interval in 2005. Same as Fig-23, scenario-based forecast in Fig-24 points out similar prediction.

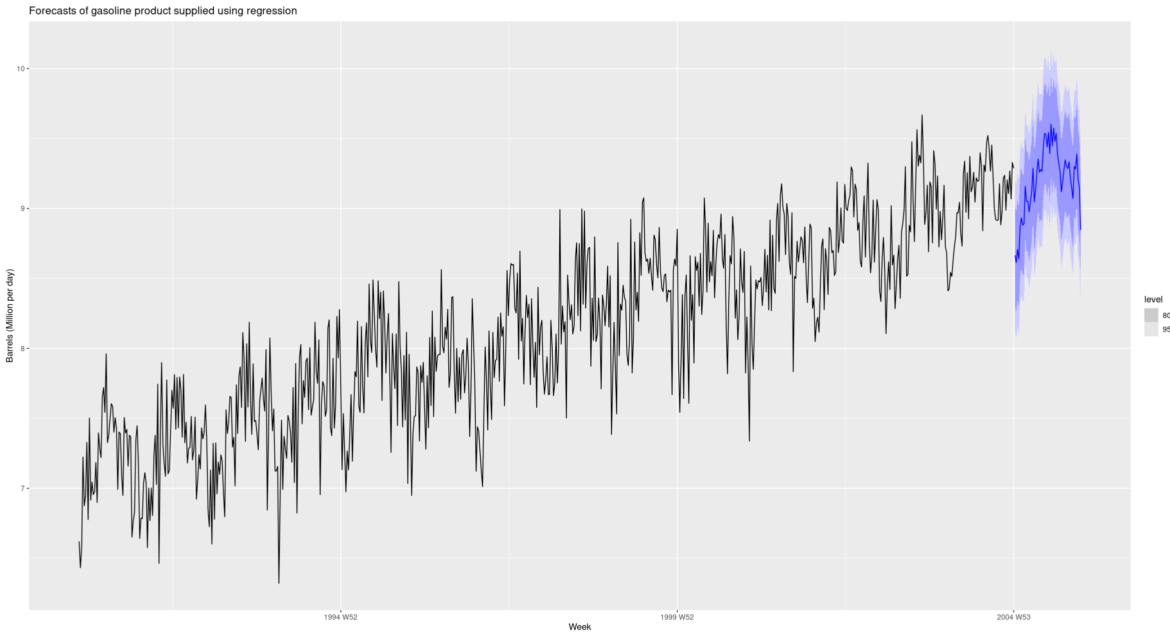

Figure 23: TSLM model forecasts for supplies of US finished motor gasoline product.

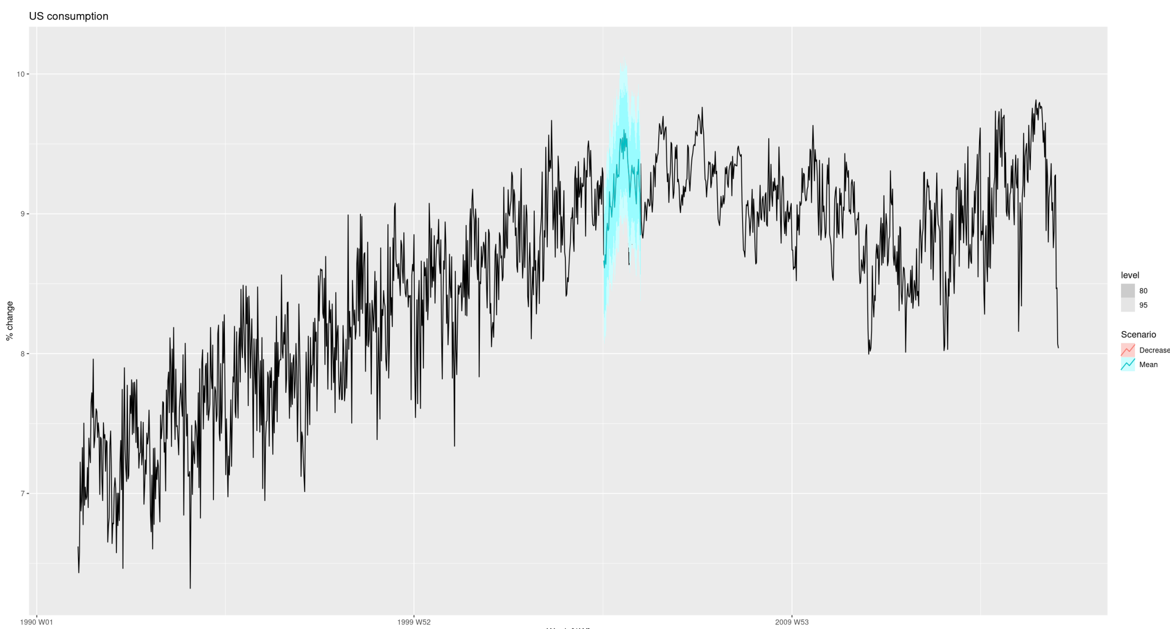

Figure 24: Scenario-based forecasts for supplies of US finished motor gasoline product.

Problem 3

Although this question is related to the population of Afghanistan, we plotted all parameters to see effect of Afghan-Soviet war on a better resolution. As we can see observations on different parameters, except population, couldn't be performed due to war between 1979-1989. We also see it in population data, during the war, population has a downward trend and decreases.

Part a

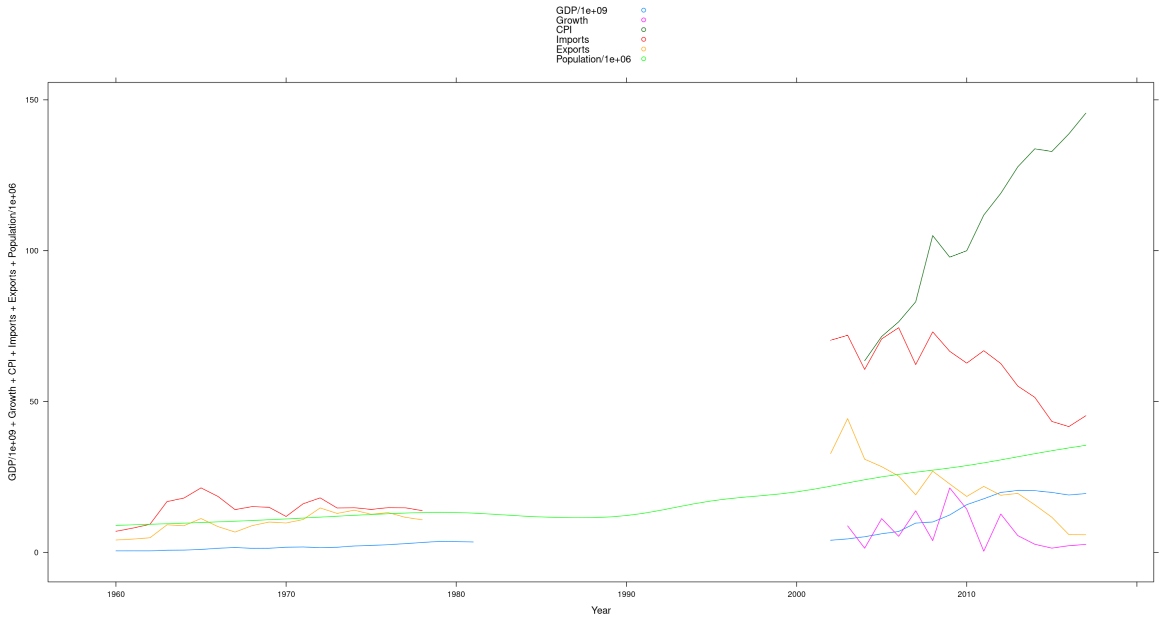

Figure 25: The annual population of Afghanistan (green line) on the data set. Other parameters are GDP (blue), CPI (cyan), Imports (red) and Exports (pink). There are no observation between 1979-1989 due to Soviet-Afghan war except population.

0.1 Part b

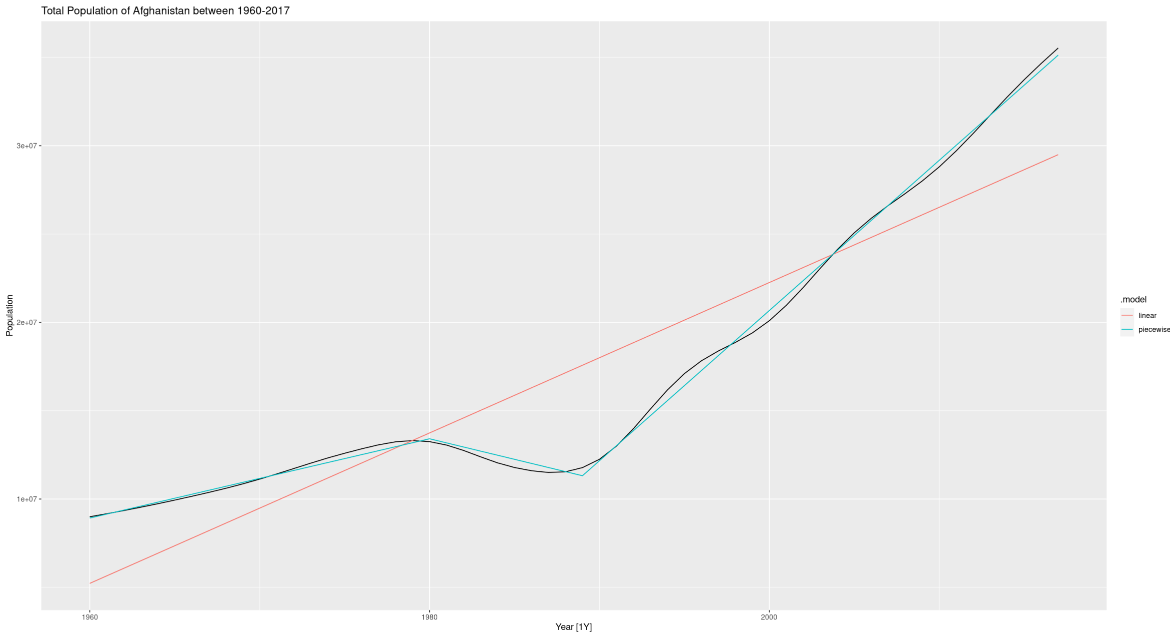

Figure 26: Linear and piecewise trend models for the population of Afghanistan data set. It is obvious that piecewise fit does better job and describe the data better than the linear fit. It is no surprise that the reason that is the pre-provided positions of knots where the historical event alters the shape of the trend.

0.2 Part c

Fig-27 demonstrate the performance difference between forecast methods undoubtedly. the perdiction of linear does not even catch the real data while piecewise method provides it with a very narrow and precise interval.

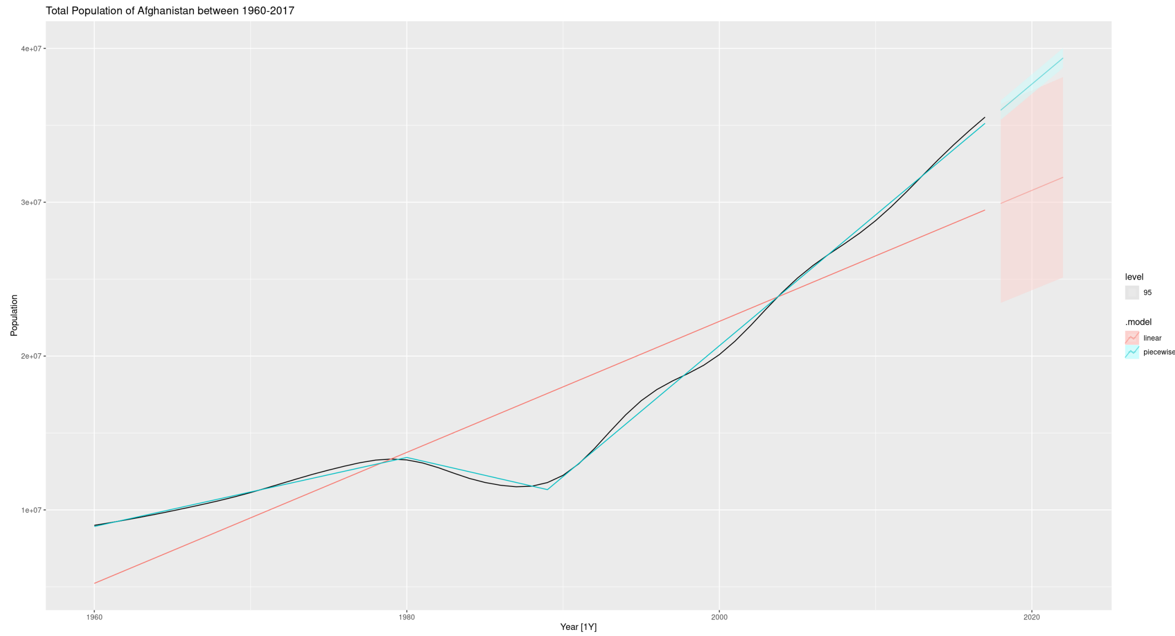

Figure 27: Forecasts with Linear and piecewise trend models for next 5 years of the Afghanistan population data set.