Final Project: Forecasting in Economics and Finance

Name: Arun Kumar Halihal Shanmukachary

Data series: "Advance Retail Sales: Retail (Excluding Food Services)"; later in the analysis, this time series will be regressed with personal income per capita using the co-integration model.

Source of Data: FRED

Objective:

To analyse and predict the discretionary retail spending of the consumers as this data series ignores the staple food spending. The data series is not seasonally adjusted and the sales figures are absolute sales figures in the dollar amount in millions. I will also use monthly personal income per capita data series to see if both the data series - income and retail sales- are co-integrated and if the personal income per capita can be used to predict the retail sales.

Observations:

As per visualization of the data series on chart, there is linear trend with slight curvature in the data pattern. It shows that the data series is likely to be non-stationary and need to be tested for random walk. If the data proves out to be non-stationary using dickey fuller test, then we need to first difference it and then test again; if the test proves differenced data to be stationary, we can then use different models such as ARIMA, Naïve, Seasonal Naïve and Moving average models for best fit. We can then test for co-integration and try to predict the retail sales using the income.

Models	Non Differenced - Non Log Data	RMSE	MAE	MPE	MAPE	MASE
Cubic Trend	Training	23880.35	17405.63	-0.6095818	5.979788	1.233463
	Test	31458.58	25094.66	0.5628037	5.411022	1.778352
Cubic Trend with Seasonal Filter/Control	Training	13521.42	10245.8	-0.1857376	3.569894	0.7260765
	Test	22861.71	20444.37	3.7392822	4.354301	1.4488056
Naïve	Training	31884.08	21070.41	-0.1925834	7.250151	1.493171
	Test	30307.61	23604.58	0.9546523	5.042142	1.672756
Rolling Model	Test	40435.75	29586.5	-0.7589479	6.509733	-0.3745058
ETS/Trend	Training	23216.43	16385.36	0.3671174	5.614951	1.161161
	Test	31862.83	24262.17	2.2538592	5.116603	1.719357
ETS/TREND+Seasonal	Training	6710.062	5285.832	0.09818049	1.916301	0.3745844
	Test	**17054.961**	15490.364	2.27371497	3.393867	1.0977361
ARIMA(2,0,0)	Test	24110.65	17527.73	-0.7096908	5.994387	1.242116
ARIMA(2,1,0)	Test	26282.29	17957.61	-0.5308114	6.232984	1.27258
ARIMA(2,1,2)	Test	21927.77	15796.82	-0.3940743	5.423311	1.119453
ARIMA(2,0,0); seasonal (1,0,0)	Test	**6835.085**	5238.672	-0.0138226	1.832555	0.3712424
ARIMA(2,1,0); Seasonal (1,0,0)	Test	**6736.754**	5001.616	-0.018622	1.749546	0.3544433
Auto.ARIMA; d=0; seasonal=true	Test	**5991.7**	4515.794	-0.0369783	1.537963	0.3200151
Auto.ARIMA; d=1; seasonal=False	Test	21927.77	15796.82	-0.3940743	5.423311	1.119453
Cointegration with error correction	Training	24540.68	17886.82	99.69828	310.2721	0.8507489
	Test	33628.11	25282.84	77.37131	104.8959	NA
Cointegration without error correction	Training	29367.09	20918.92	128.5503	292.0381	0.9949644
	Test	38987.45	30135.56	41.36767	112.5477	NA

Models	Differenced Log Data	RMSE	MAE	MPE	MAPE	MASE
Cubic Trend	Training	0.02692903	0.02129032	101.406	103.0732	0.6832644
	Test	0.04191207	0.02826114	114.8106	114.8106	0.9069772
Naïve	Training	0.04558316	0.03612286	170.6503	429.9599	1.1592812
	Test	0.04191914	0.0280449	124.3245	124.3245	0.9000374
Rolling Model	Test	NA	NA	NA	NA	NA
ETS/Exponential	Training	0.02947077	0.02345029	96.21012	136.8449	0.7525838
	Test	**0.04189722**	0.02789586	130.96311	130.9631	0.8952544
ETS Auto/Exponential	Training	0.0269355	0.02127249	100.7098	100.7098	0.6826922
	Test	0.04214921	0.02857223	100.8354	100.8354	0.9169609

Below are the observations made after performing the time series regression analysis. The remarks are here and the respective figures/charts are posted in the following pages.

Fig 1: Normal plot (plot function): Indicates there is trend which is slightly non-linear, has seasonal pattern where there is a sudden surge in sales in December and then a slump in sales in January.

Fig 2: Normal plot (using autoplot function): As above plot, indicates there is trend which is slightly non-linear, has seasonal pattern which is sudden surge in sales in December and then a slump in sales in January.

Fig 5: Seasonal Chart: The chart below shows clear seasonal trend where sales surge during the month of December and then drop in January.

Fig: 6 Autocorrelation Plot: Using the p <- ggAcf (retail.ts), the chart was derived which shows the correlation between the lags. The plot shows that there is still strong correlation that exits between data in time T (lag 0) and lag 25. So the time series is not just dependent on the previous lag but also on the lagged data from much previous period as far as lag 25. There is also seasonal trend that is exhibited by point lag 12 and lag 24.

Fig : 9 Plotting the ACF of residuals of Cubic Trend model using ggAcf: The cubic trend model fits the retail sales data well as the correlation of residuals of fitted model are well below the 0.3. However, since the seasonal component of the model is not controlled, the lag at 12, 24, 36…etc are strongly correlated.

Fig 10) Plotting the ACF of residuals of Cubic Trend model with differenced log data using ggAcf: The cubic trend model on the differenced log data fits the retail sales data well as the correlation of residuals of fitted model are well below the 0.3 and more consistent and random. Also, the seasonal component of the model is controlled, so the lag at 12, 24, 36…etc are not strongly correlated and behave similar to other lags.

Fig 21:) ACF of Trend model controlled for seasonality: Since the data series behaves as a random walk but controlled for seasonality, we can see that the ACF of residuals are strongly correlated even for lags further the sixth period. But the autocorrelation at lag12 looks much similar to other lags since seasonality is controlled.

Fig: 22) ACF of ARIMA (2,0,0): Since the ARIMA model used stationary data series, the residuals of regression look less correlated at any lag. Since the autocorrelation stays below 0.3 at any lag, ARIMA (2,0,0) model fits the data series well. However the seasonality is not controlled for, hence we can see lag12 and lag24 exhibit high autocorrelation.

Fig 23) ACF of AutoArima Model --> train.ts, d=0, ic="bic",seasonal=TRUE--> ARIMA (3,0,1) (1,1,1) [12] with drift: Since the ARIMA model used stationary data series, the residuals look less correlated at any lag. Since the autocorrelation stays below 0.15 at any lag, ARIMA (2,0,0) model fits the data series well. Since, the seasonality is controlled for, we can see lag12 and lag24 exhibit very low autocorrelation.

Dickey Fuller Results: Here t-statistics of 78.34 > critical value of 6.3, which implies rejection of the null hypothesis of random walk. So as per dickey fuller test, the data series is stationary even though it looks like a random walk when visually inspected (as it exhibited a trend).

```
Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
   Min     1Q Median     3Q    Max
-66167 -12822   -668  11780  73810

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.377e+05  1.121e+04  12.284   <2e-16 ***
z.lag.1     -9.080e-01  7.254e-02 -12.517   <2e-16 ***
tt           7.876e+02  6.467e+01  12.178   <2e-16 ***
z.diff.lag   6.803e-02  5.597e-02   1.216    0.225
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24290 on 321 degrees of freedom
Multiple R-squared:  0.4281,	Adjusted R-squared:  0.4227
F-statistic: 80.09 on 3 and 321 DF,  p-value: < 2.2e-16


Value of test-statistic is: -12.5168 52.5352 78.3425

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

Results of Co-integration between retail sales and Per capital income with error correction: P value of less than 0.05 proves that we reject the null hypothesis and state that the residuals of the cointegration are stationary which proves that retail sales and per capita income are co-integrated.

```
> print(summary(ec.model))

Call:
lm(formula = ret.ts ~ retlag.ts + PINLAG.ts + creglag.ts, data = veclags.ts)

Residuals:
   Min     1Q Median     3Q    Max
-71540 -13667   1100  11700  75446

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 2008.22897 1551.71585   1.294    0.196
retlag.ts      0.03791    0.05641   0.672    0.502
PINLAG.ts    -23.53182   18.29875  -1.286    0.199
creglag.ts    20.55332    1.72650  11.905   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25000 on 331 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.4099,	Adjusted R-squared:  0.4046
F-statistic: 76.65 on 3 and 331 DF,  p-value: < 2.2e-16
```

Results of Co-integration between retail sales and Per capital income without error correction: P value of less than 0.05 proves that we reject the null hypothesis and state that the residuals of the cointegration are stationary which proves that retail sales and per capita income are co-integrated. Note that P value is slightly higher compared to the cointegration with error correction model above.

```
> print(summary(ecbench.model))

Call:
lm(formula = ret.ts ~ retlag.ts + PINLAG.ts, data = veclags.ts)

Residuals:
   Min     1Q Median     3Q    Max 
-88443 -11651   2833  12906  73610 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2397.33651 1851.17943   1.295    0.196    
retlag.ts     -0.38724    0.05211  -7.432 9.17e-13 ***
PINLAG.ts    -26.27929   21.83331  -1.204    0.230    

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29830 on 332 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.1573,	Adjusted R-squared:  0.1522 
F-statistic: 30.98 on 2 and 332 DF,  p-value: 4.607e-13

>
```

Final Results of Analysis:

- The metrics RMSE (root mean squared error), MAPE (Mean Absolute Percentage Error) and MAE (Mean Absolute Error) were used to decide on the performance of the models.
- The ARIMA model of the order (3,0,1) (1,1,1) [12] with drift using Auto Arima function with d=0 and seasonal = true makes the best fit with least RMSE (root mean squared error) of 5991.7. The model also has least MAE (Mean Absolute Error) of 4515, MAPE (Mean Absolute Percentage Error) of 1.53.
- The other ARIMA models with seasonality controlled performed better compared to cubic trend, Naïve, ETS, co-integration model and “ARIMA models where seasonality was not controlled” when looked at RMSE and other error measures.
- Cubic trend model, Naïve model and rolling model were under performers.
- Co-integration with error control performed better than the Co-integration without error control. However both models underperformed compared to ARIMA model.
- ARIMA models with first differencing did not perform well in predicting the data series compared to ARIMA models that were controlled for seasonality which implies that the data series was indeed stationary as implied Dickey fuller test and first differencing added very less value to the model.
- Co-integration of retail sales with personal income per capita showed that personal income alone cannot predict the retail sales and there are number of other factors that drive the retail sales. This was obvious in very high RMSE.
- Models used for differenced data cannot be compared with models where raw data was used, because the output given by models that used differenced data had to be reworked back to raw output. So comparing only those models that used differenced data, ETS/exponential filter model performed better.

Different plots:

Fig1) **Normal plot (plot function)**

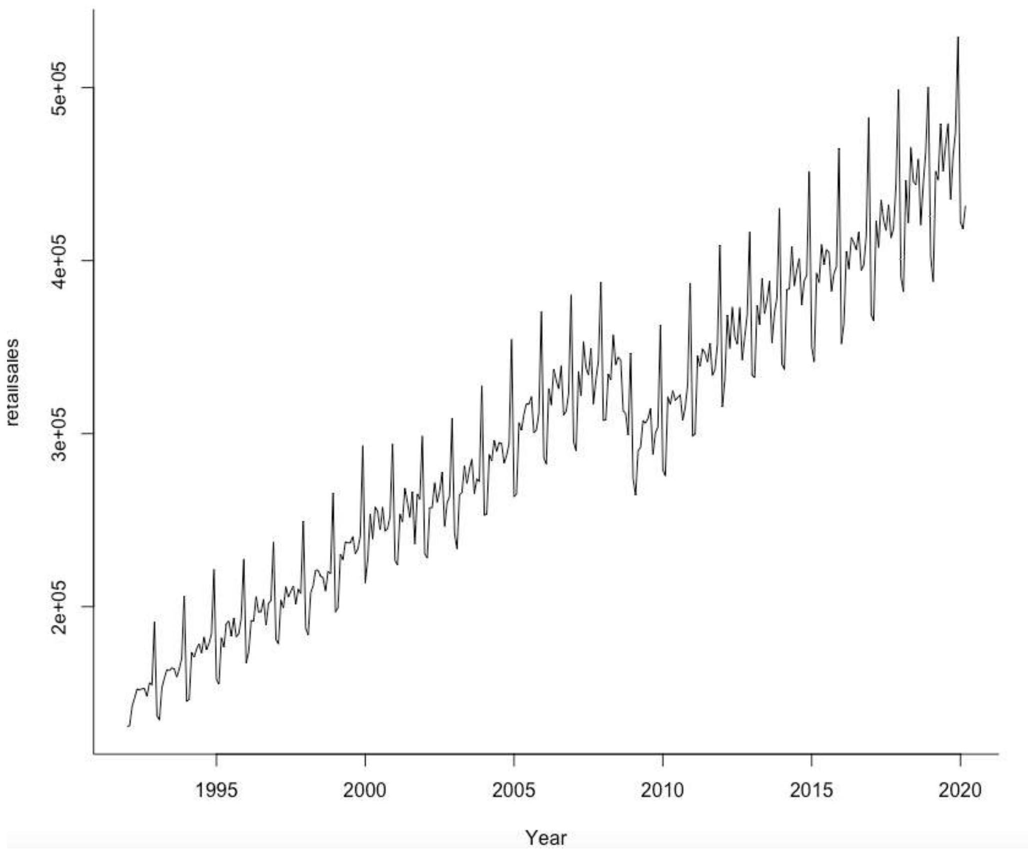


Fig : 2) **Normal plot (using autoplot function)**

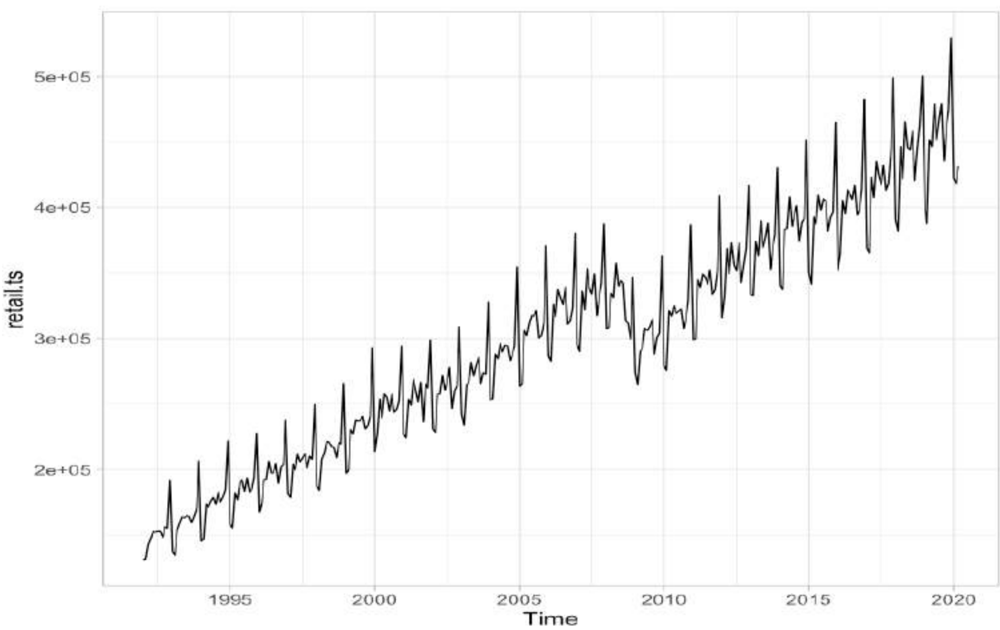


Fig 3: Linear Trend fitted model

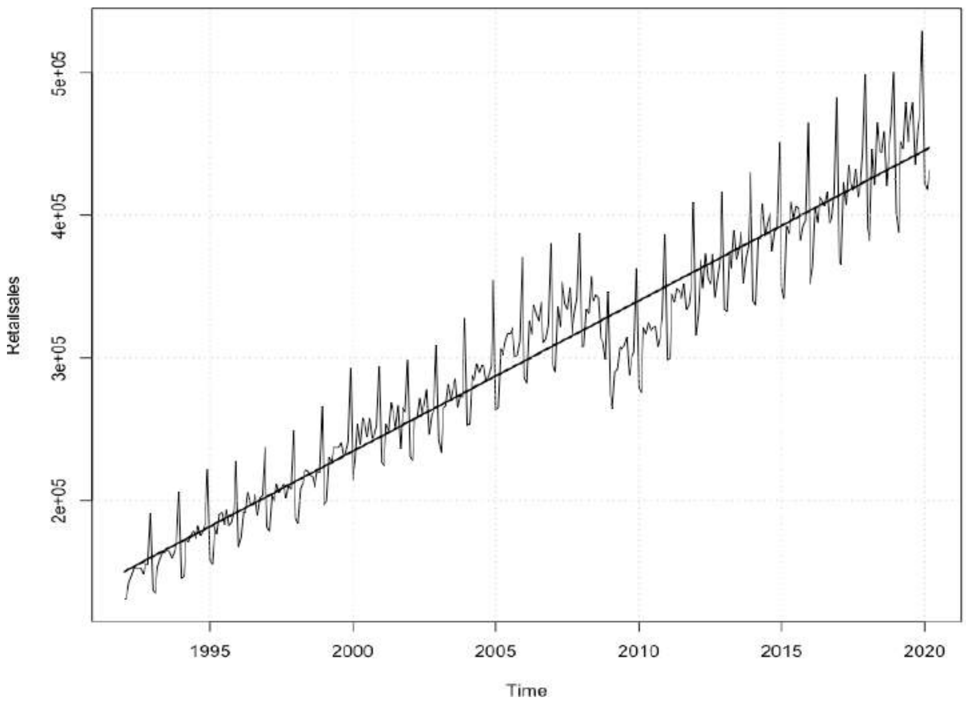


Fig:4) Cubic Trend fit model

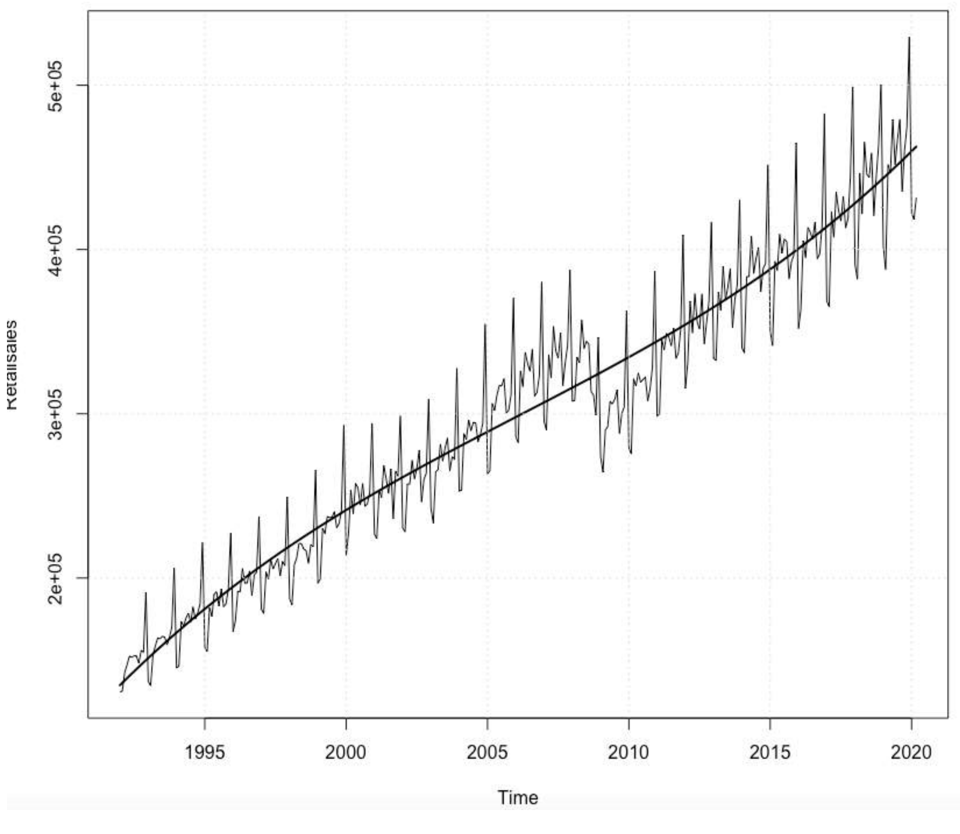
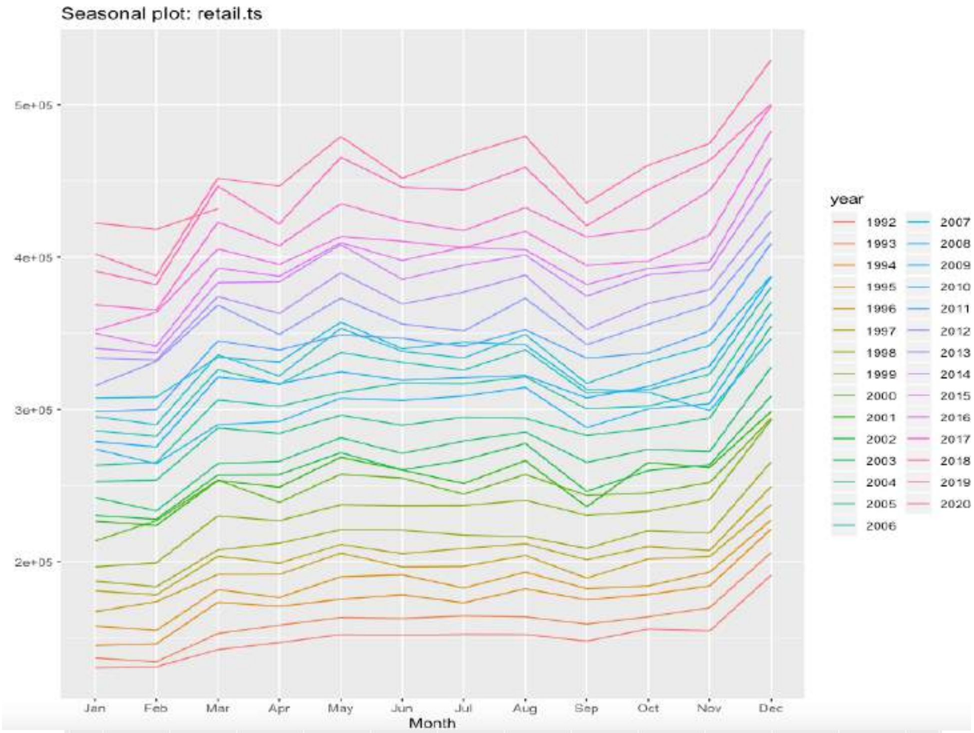
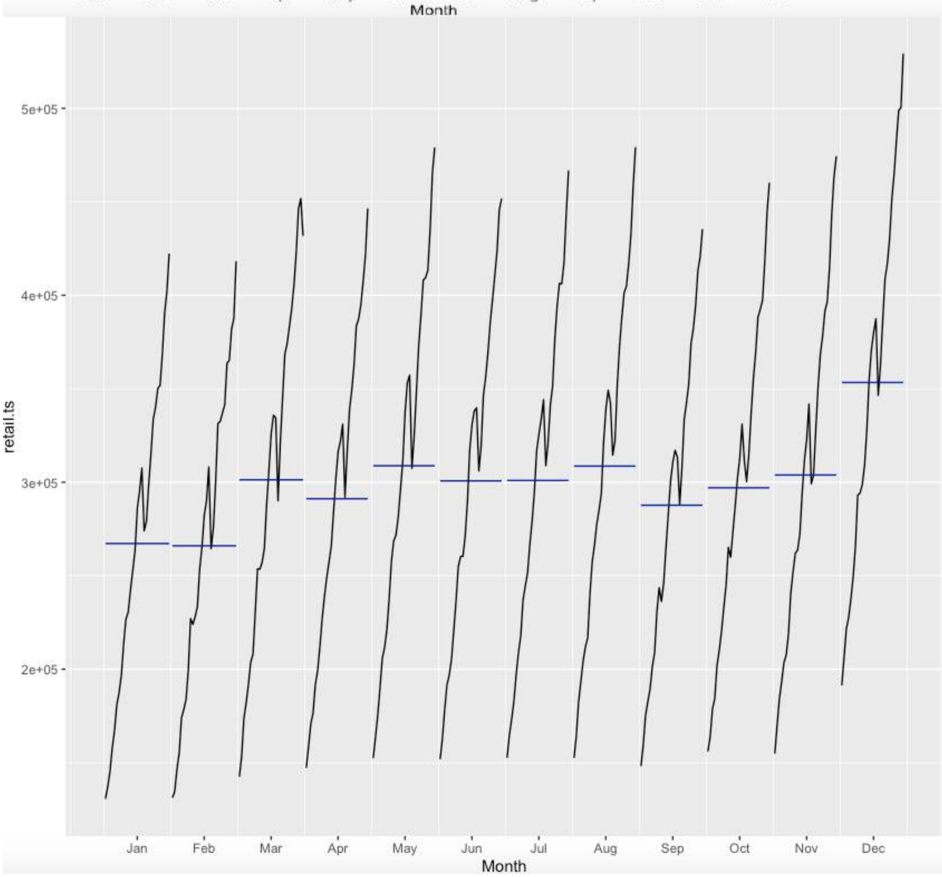


Fig: 5) Seasonal Chart


Seasonal plot: retail.ts
5e+05
4e+05
3e+05
2e+05
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
year
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020


retail.ts
5e+05
4e+05
3e+05
2e+05
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

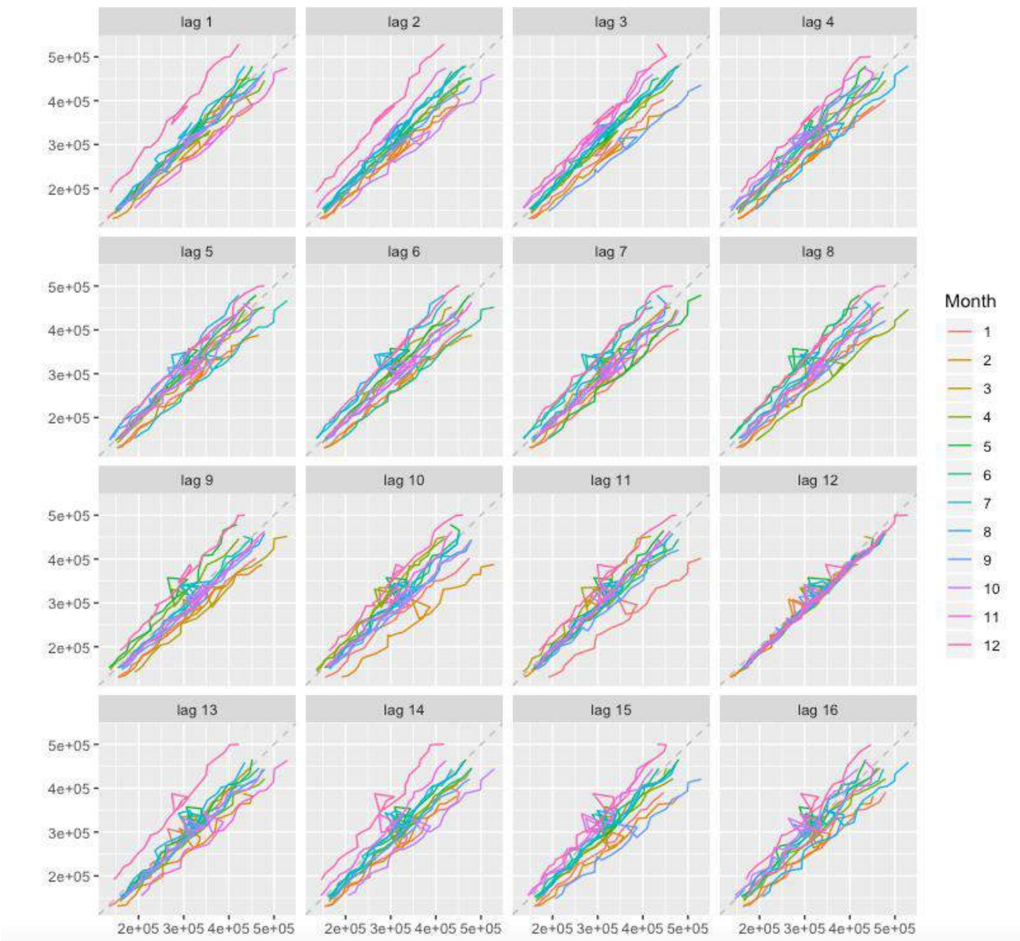


Fig: 6 Autocorrelation Plot of retail sales time series

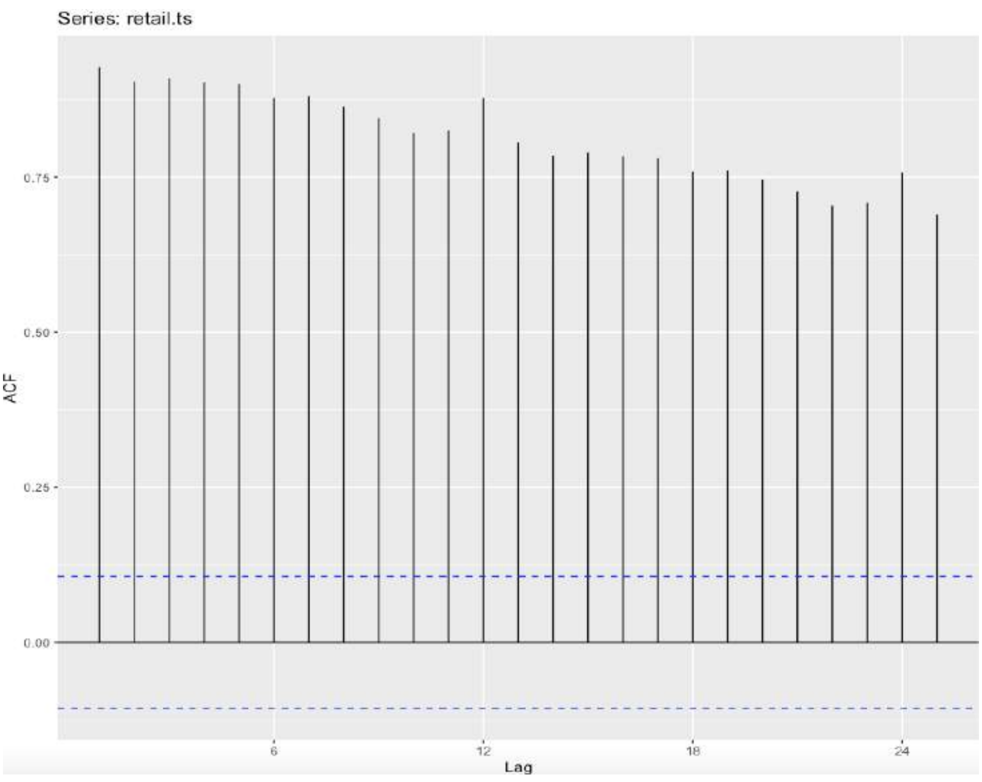


Fig 7) Cubic Trend Model Forecast Plot using non-differenced/normal data

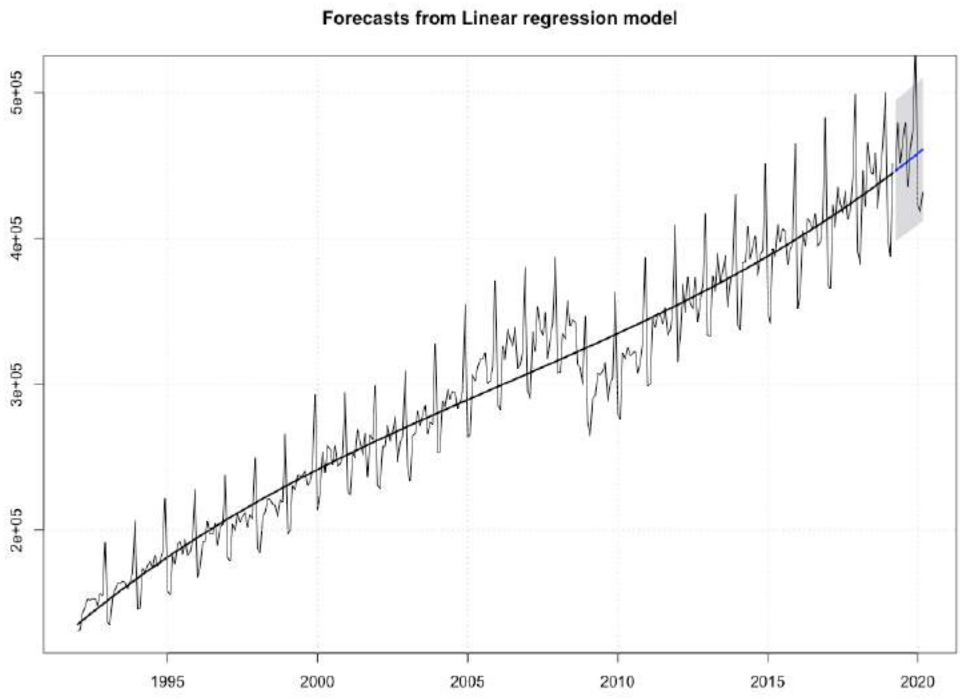


Fig 8) Cubic Trend Model Forecast Plot using differenced log data

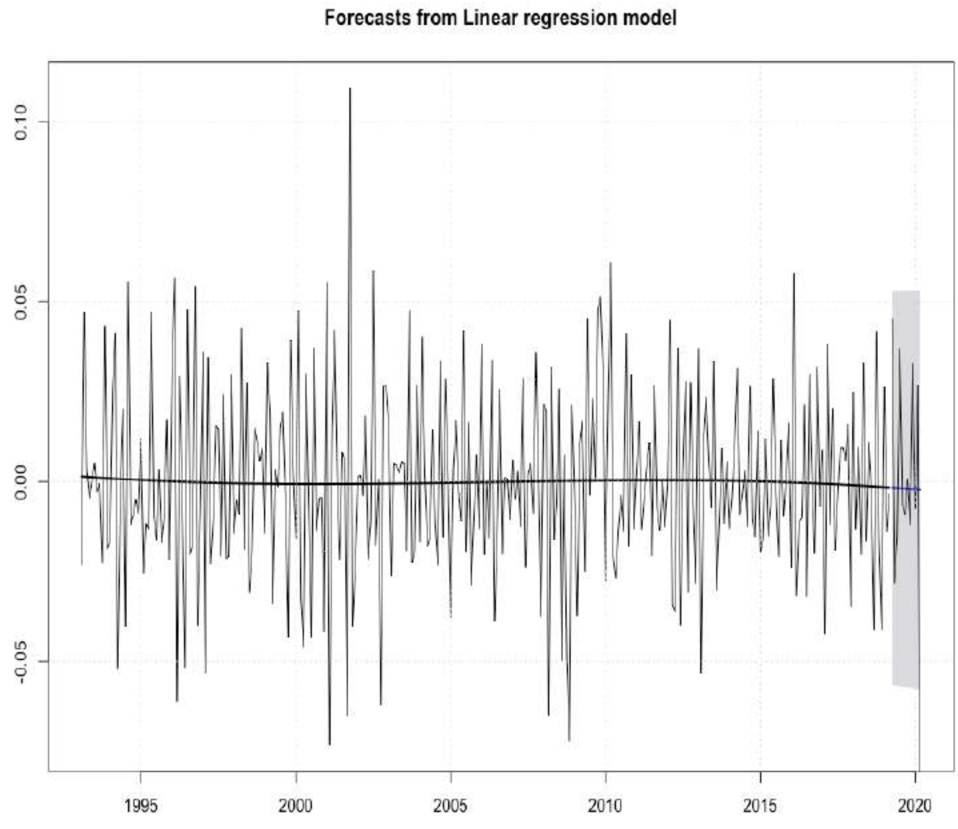


Fig 9) Ploting the ACF of residuals of Cubic Trend model using ggAcf:

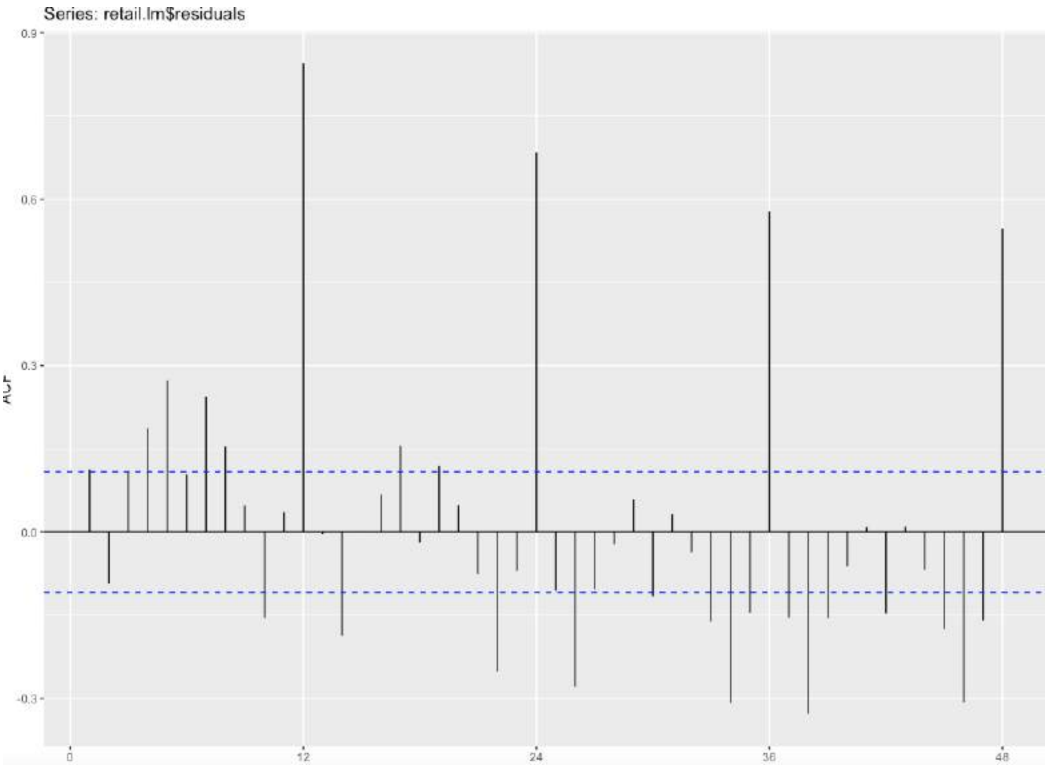


Fig 10) Ploting the ACF of residuals of Cubic Trend model with differenced log data using ggAcf

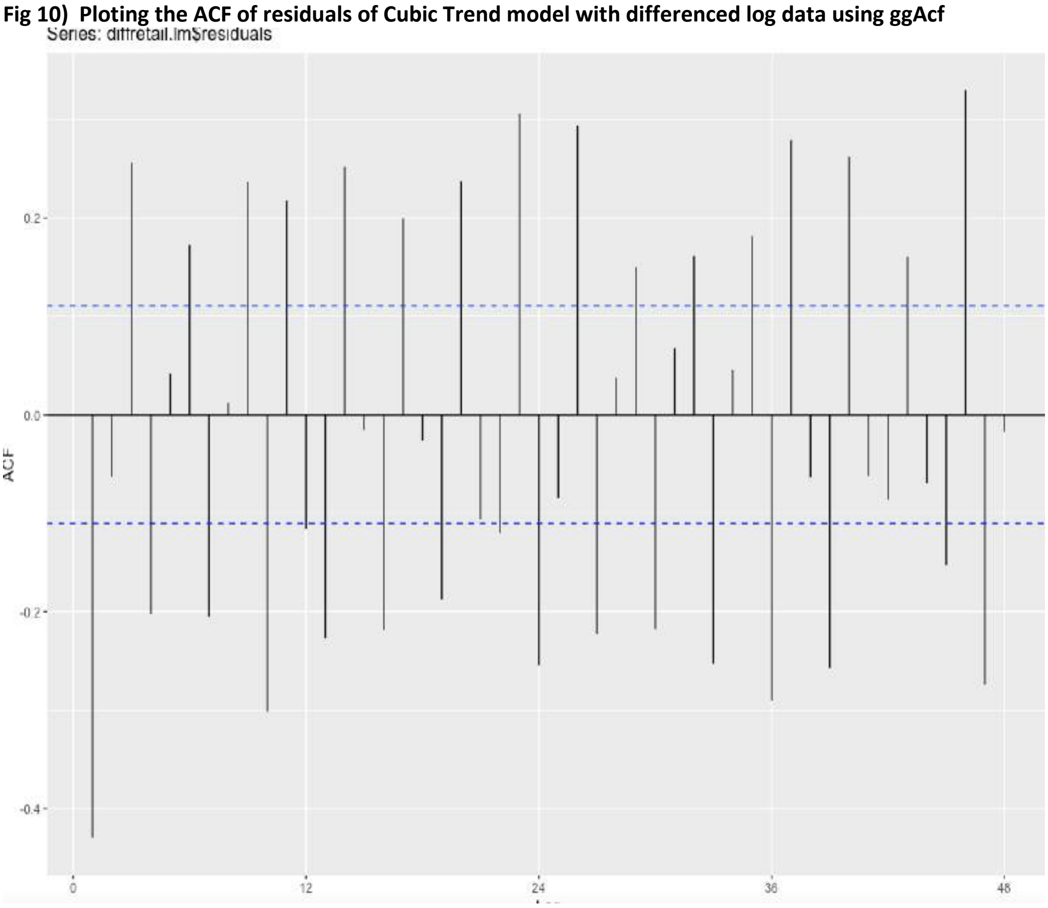


Fig 11) Cubic trend model fitted with cubic, Naïve, rolling random walk forecast

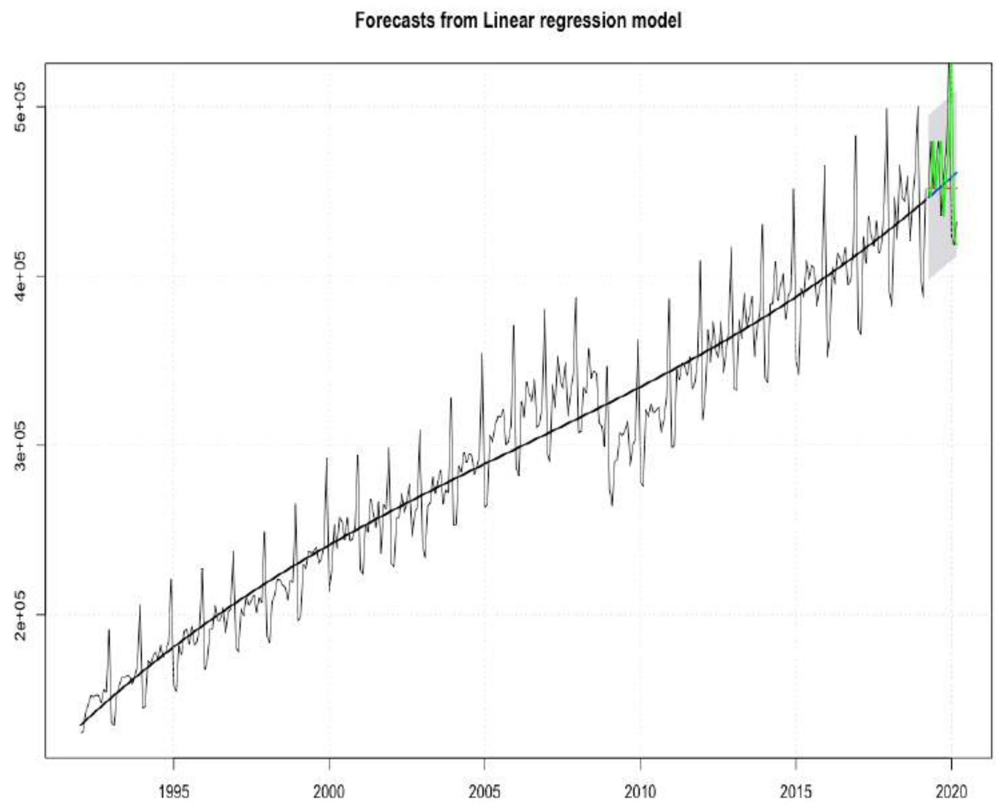


Fig 12) **Plot of different Naïve Forecast**

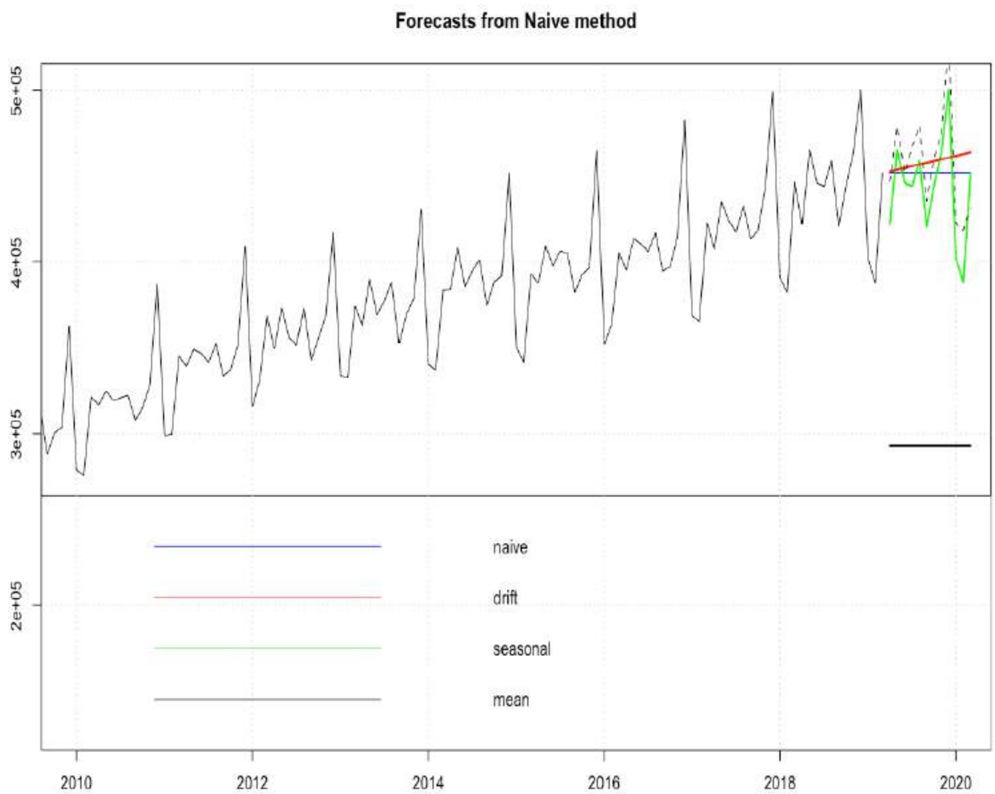


Fig 13) **Cubic trend model fitted with cubic, Naïve, rolling random walk forecast on differenced log data**

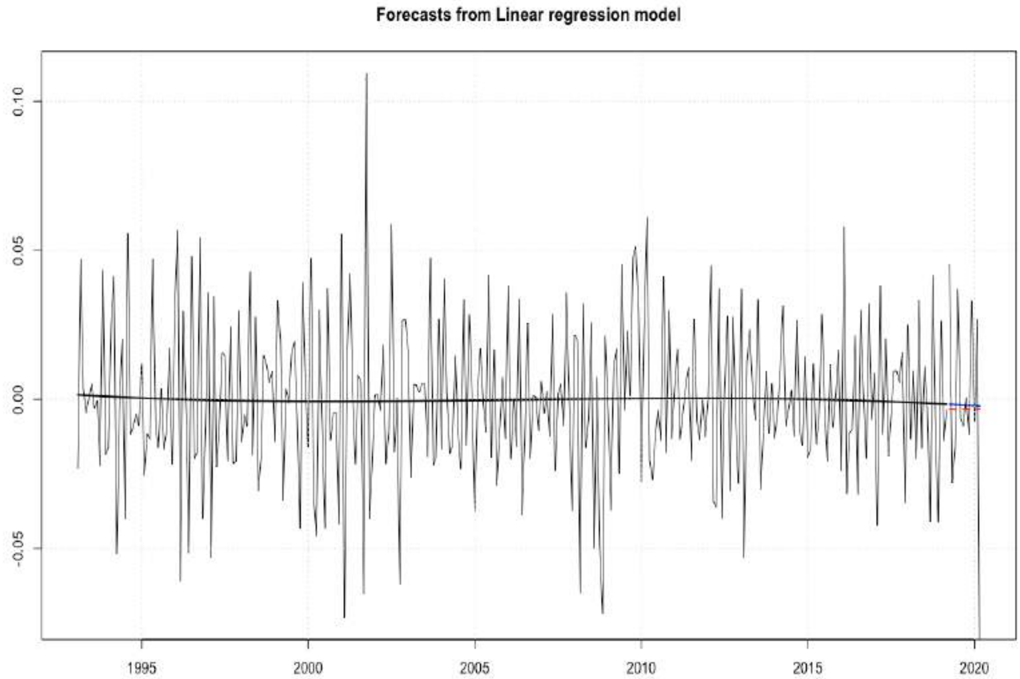


Fig 14) Plot of different Naïve Forecast using differenced log data series

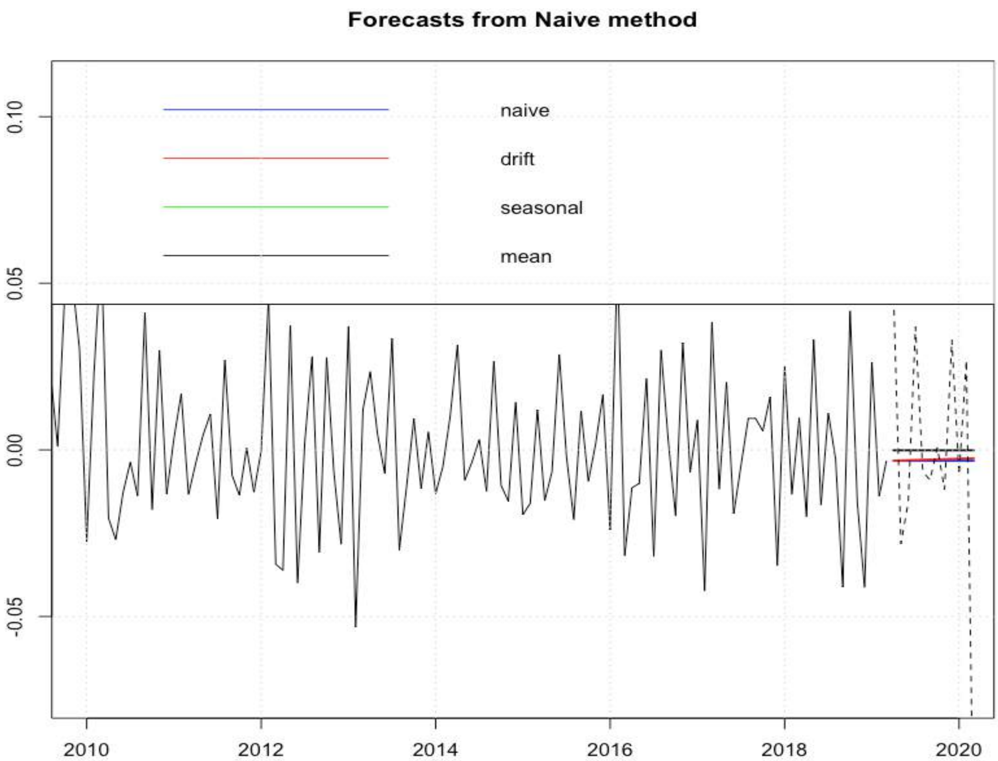


Fig 15) Moving Average model

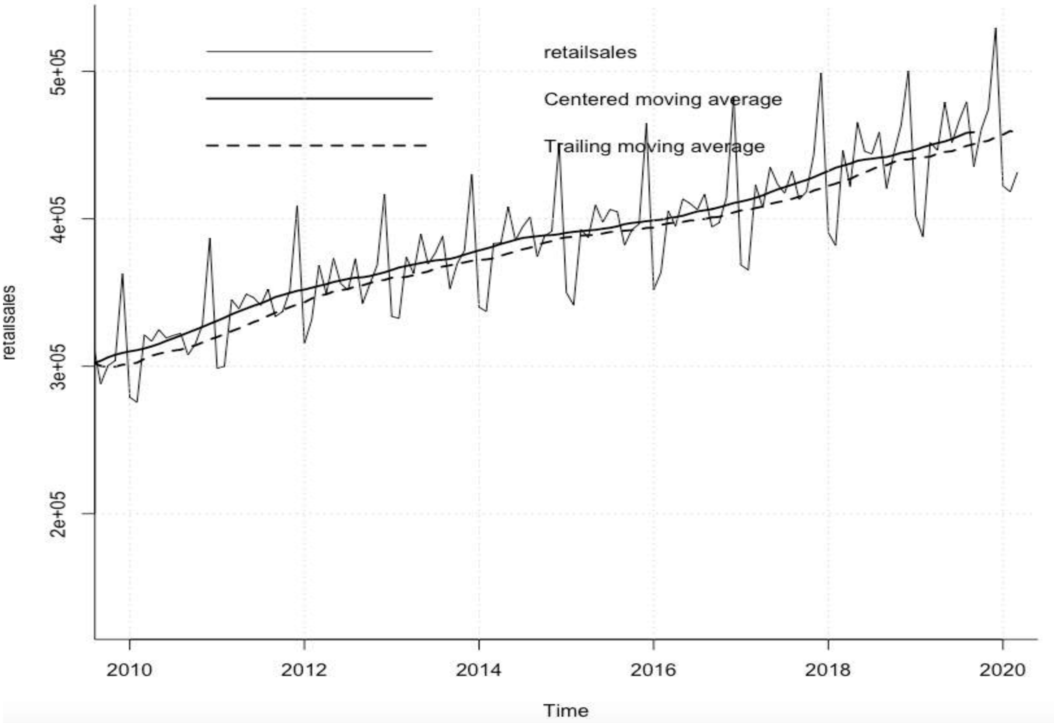


Fig 16) Moving Average model forecast plot

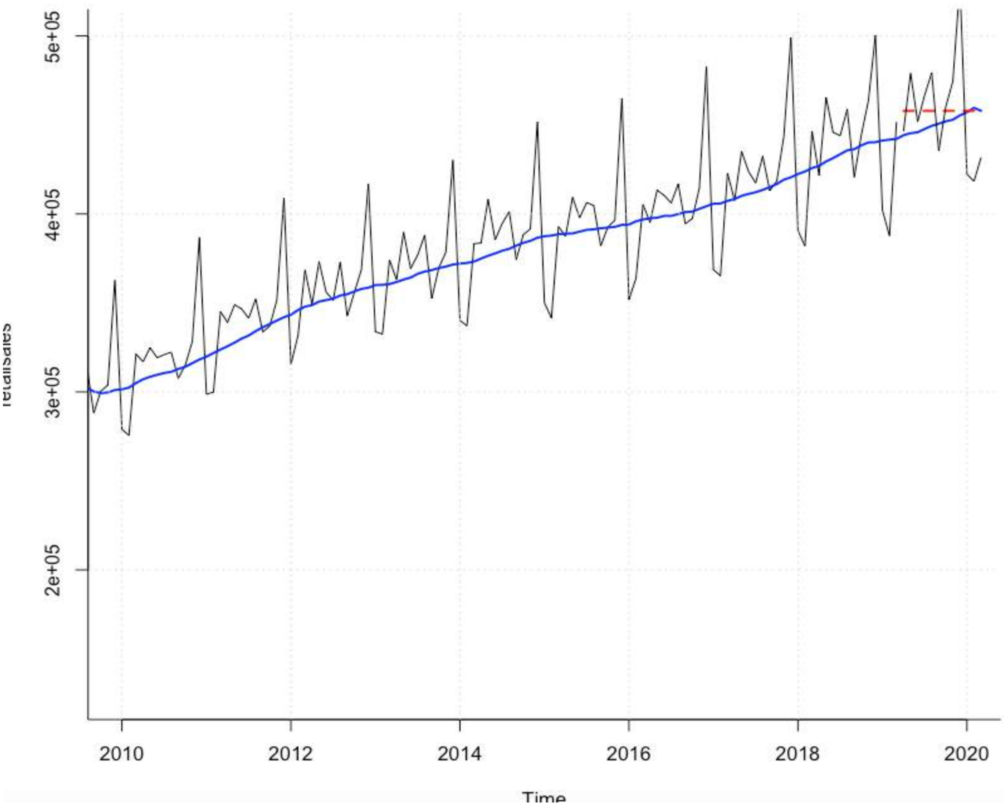


Fig 17) Exponential Filter Forecast on Differenced log Data

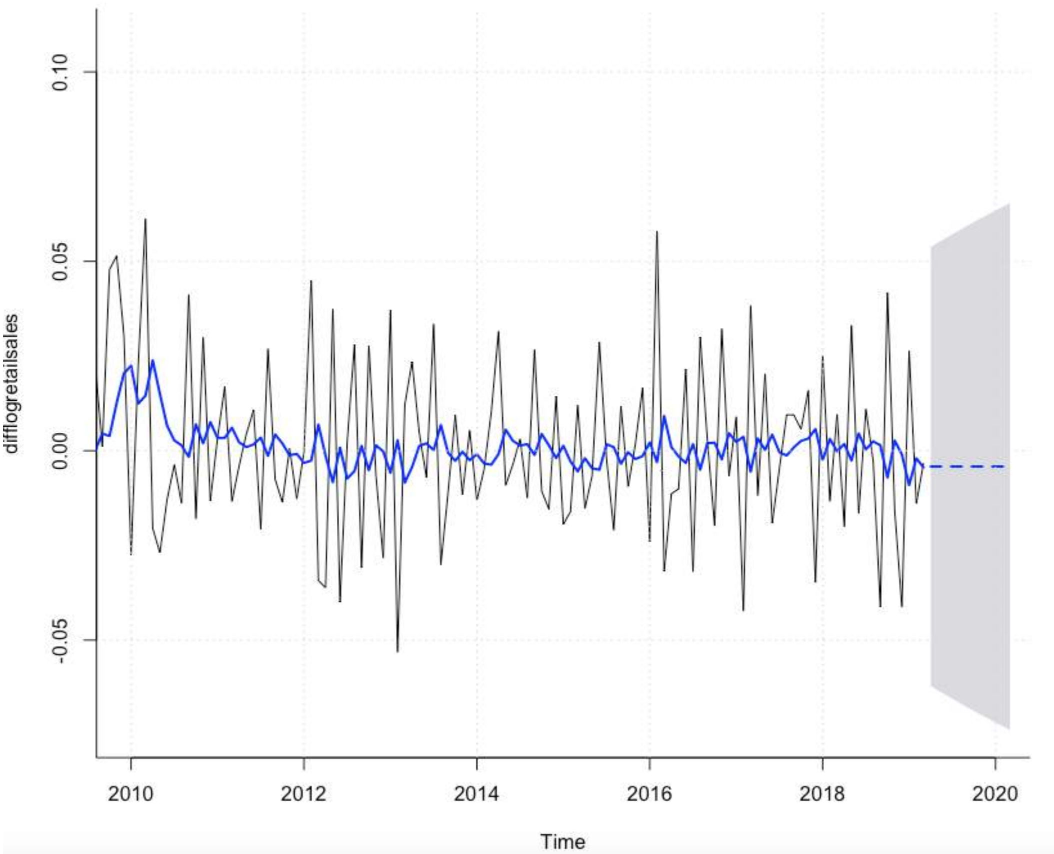


Fig 18) Plot of ETS/Exponential model capturing Trend and Seasonality using ETS Function

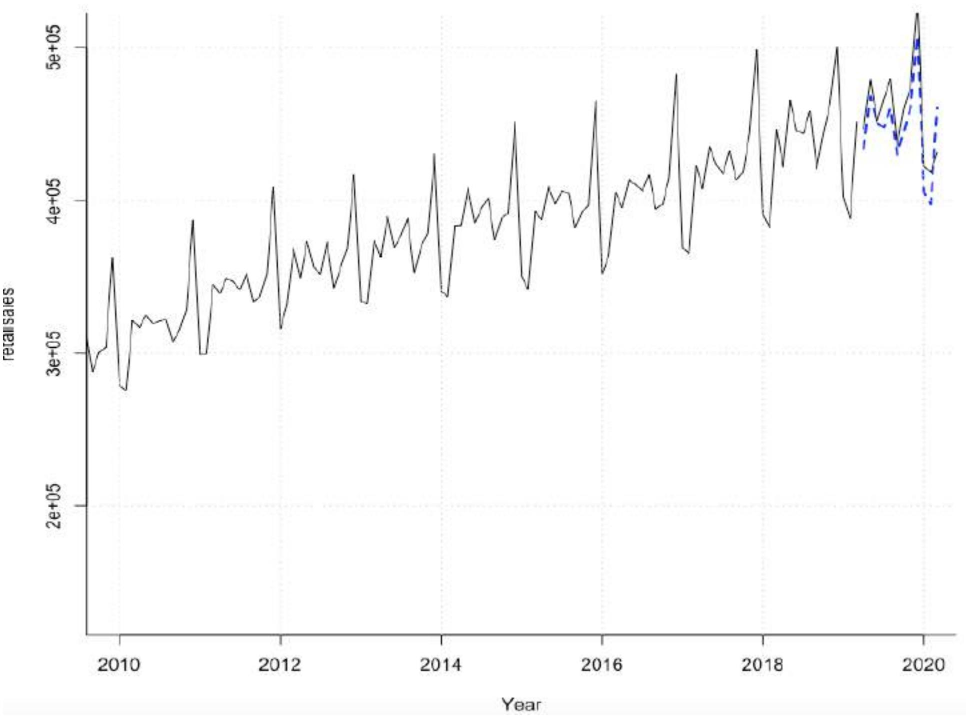


Fig 19) Forecast for AUTO.ARIMA with d=0 and Seasonal=True

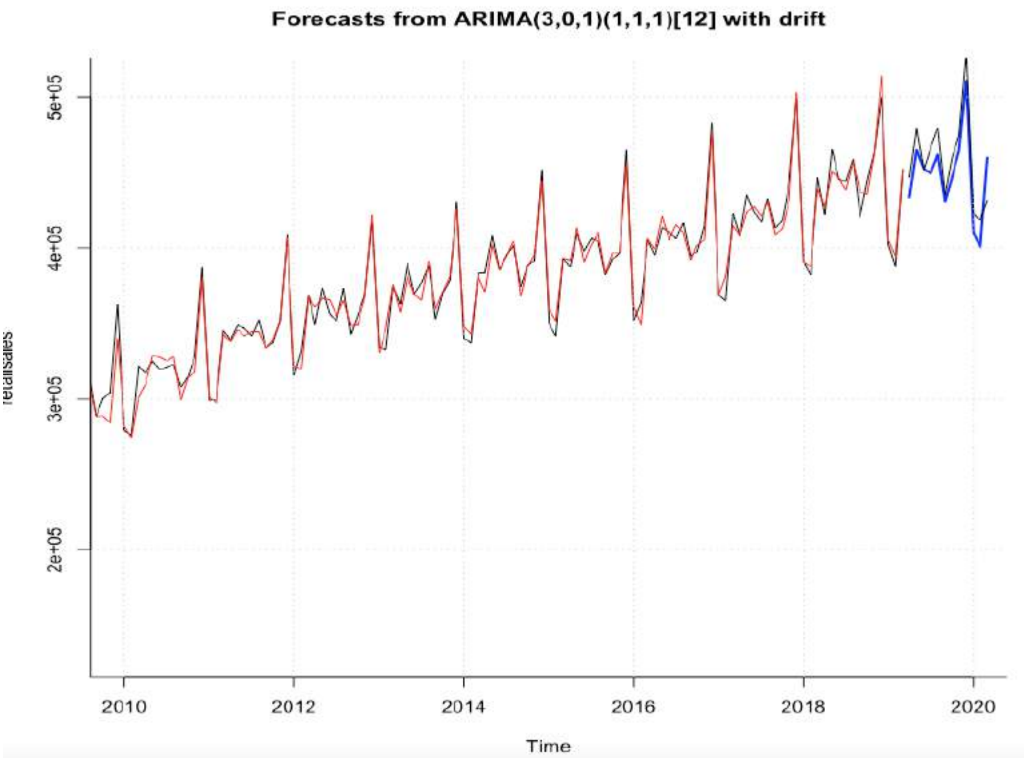


Fig 20) Forecast for Trend model controlled for Seasonality

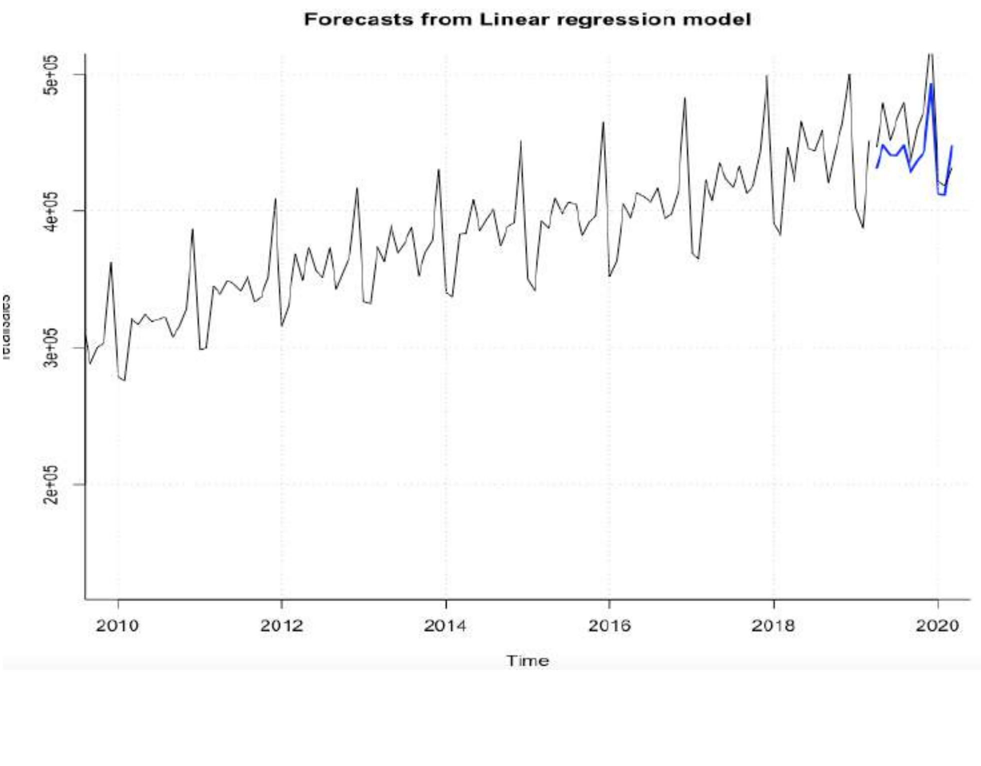


Fig 21:) ACF of Trend model controlled for seasonality:

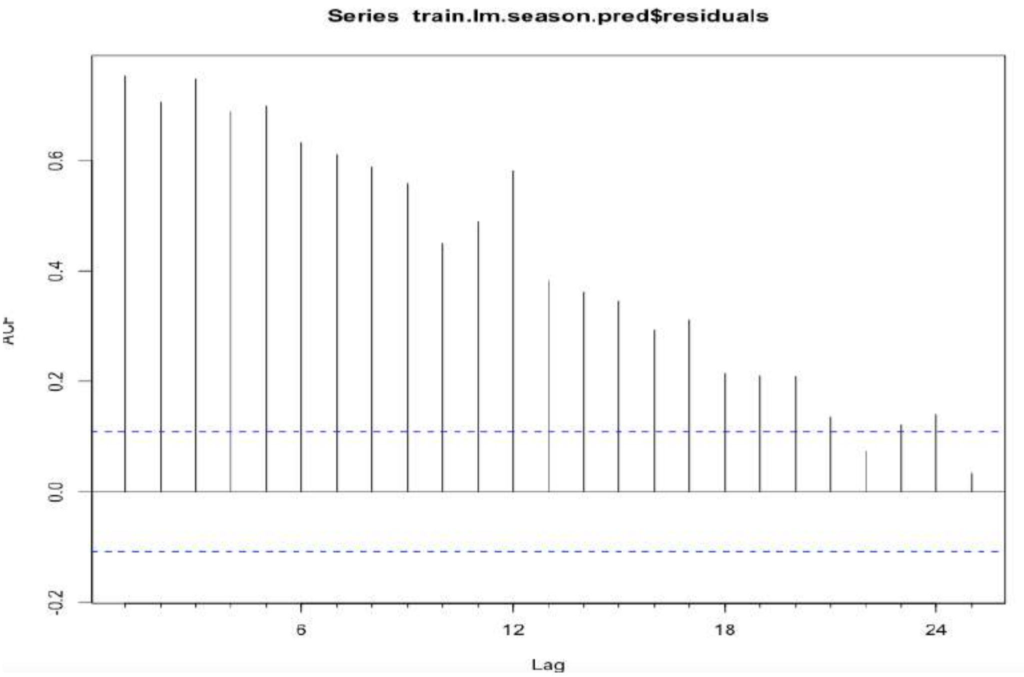


Fig: 22) ACF of ARIMA (2,0,0)

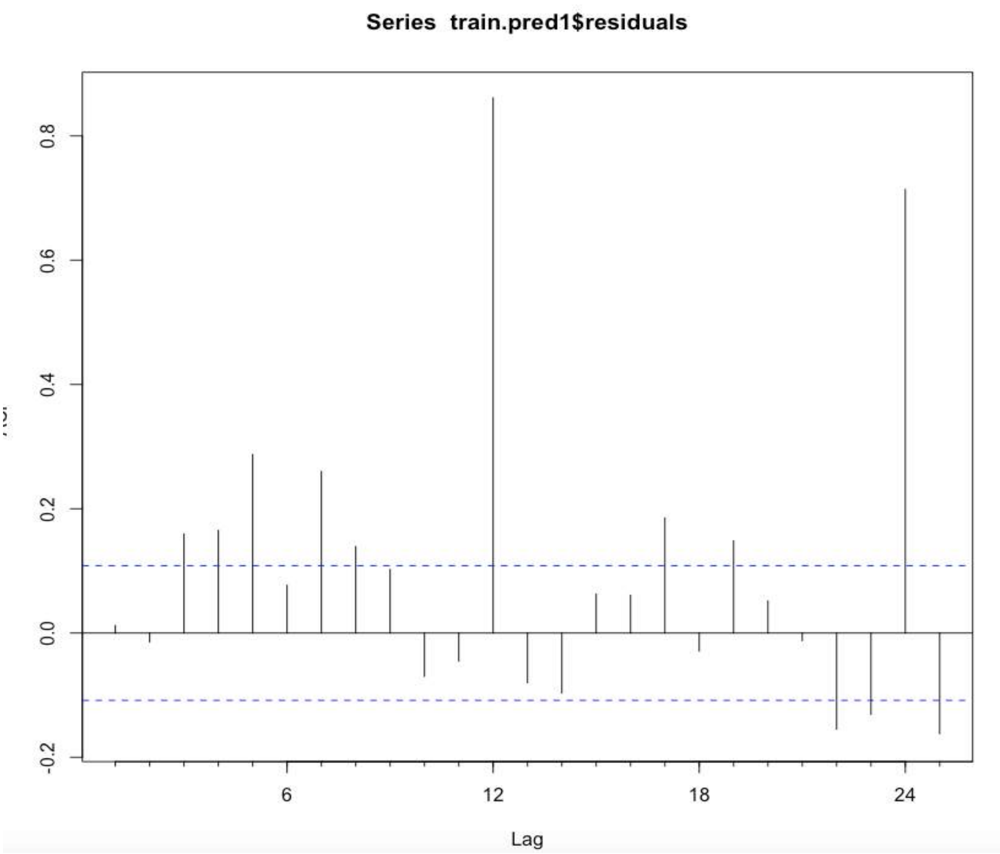
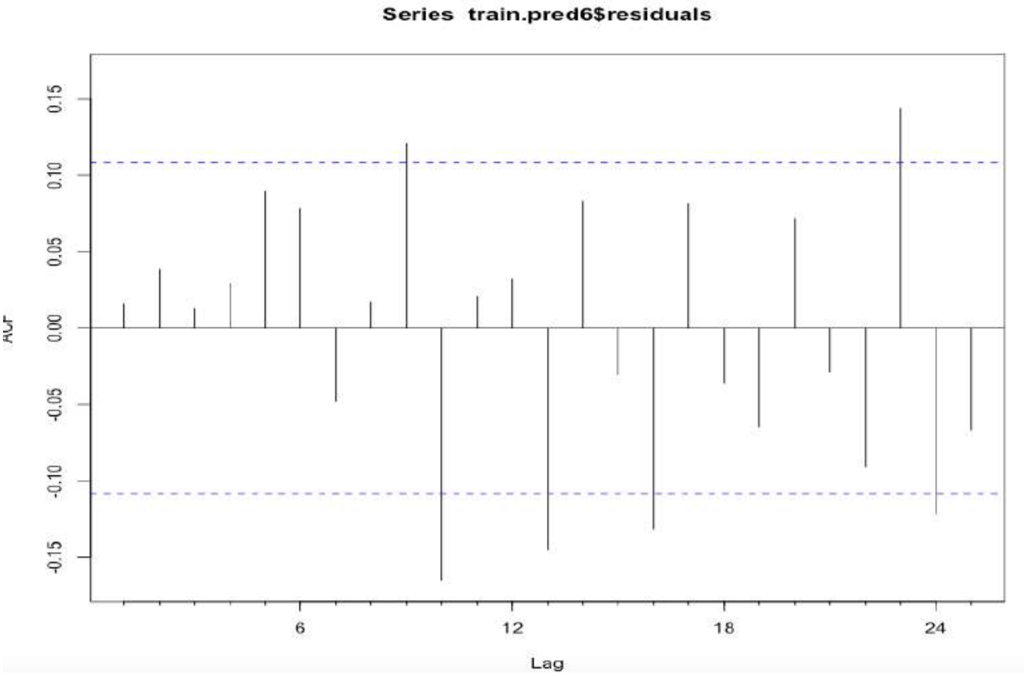


Fig 23) ACF of AutoAurima Model --> train.ts,d=0,ic="bic",seasonal=TRUE--> ARIMA (3,0,1) (1,1,1) [12] with drift



Thank You