DSC 425 Time Series Analysis
Group Project Final Report
Amalraj Dominick
Chowaniak Michal
Nguyen Long

Advance Monthly Retail and Food Services

Introduction

The Advance Monthly Sales for Retail and Food Services data is part of a bigger Monthly Retail Trade survey data, which *"provides early market indication of current retail trade activity in United States. Retail sales are one of the primary measures of consumer demand for both durable and non-durable goods"[1].* Survey is mailed to representative sample of about 5,500 firms of different sizes 5 working days before end of reporting period and responses are required 2 working days after end of reporting period. Around 1,250 firms, because of their size and large effect on sales, are always included in the sample. Advance Monthly Retail and Food Services data is used frequently by government and businesses for example it is used as input for estimating Gross Domestic Product by Bureau of Economic Analysis, and Federal Reserve Board uses it as prediction of economic trends.

Exploratory Analysis

In this section we explain preprocessing steps. At the beginning data was loaded, there was no missing values. We looked into exploratory analysis for both the adjusted and non-adjusted data. Below plots show adjusted and non-adjusted data. There is a steady growth, non-adjusted shows seasonality. There are a small shock around year 2001 and large one around year 2008. The histogram for both the adjusted and non-adjusted are skewed to the left, as there is a tail for both histograms to the right. From the decomposition plots there is no sign of seasonality with the extreme values on the y axis, even though the seasonality plot shows a similar pattern throughout. The trend shows a linear increasing trend, with an unusual shock around 2008 like mention above. The time plot is shown on figure 1, the histogram are shown in figure 2, and the decomposition plots are shown in figure 3.

[1] United States Census Bureau, "About the Advance Monthly Retail Trade Survey", March 16, 2019). Retrieved from https://www.census.gov/retail/marts/about_the_surveys.html

Figure 1

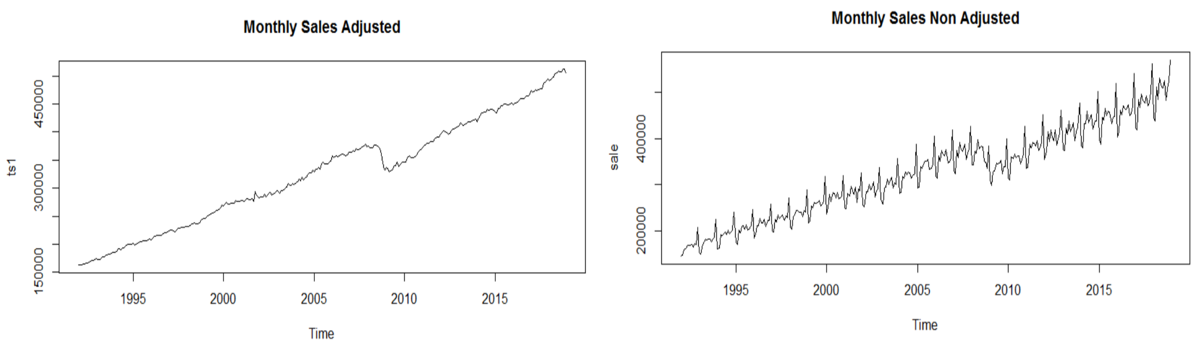


Figure 2

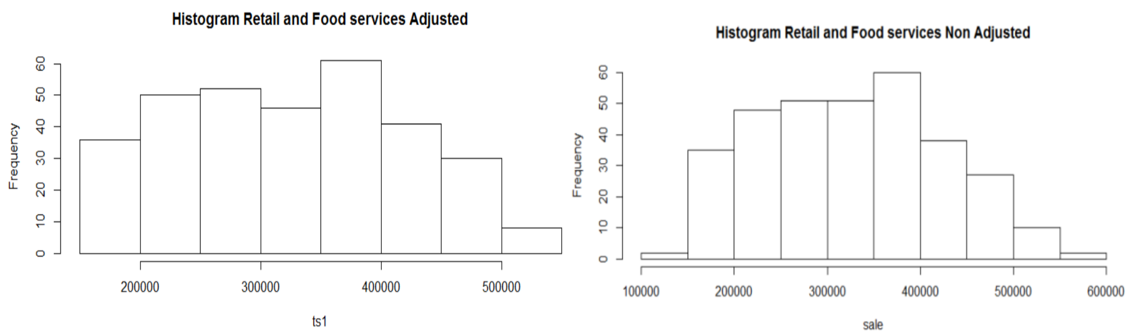
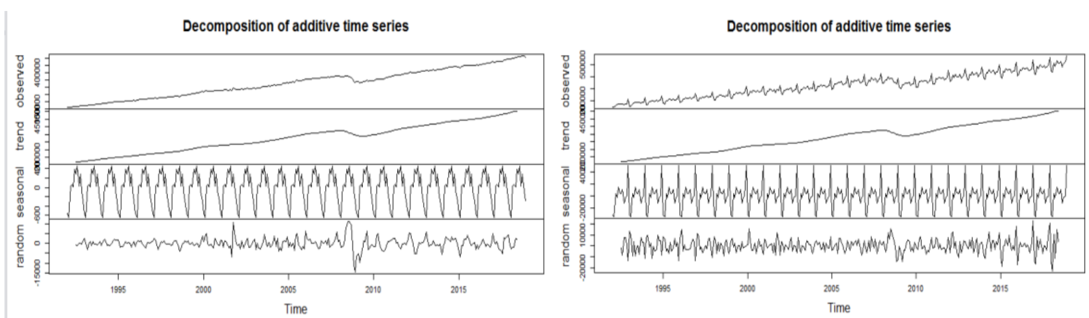
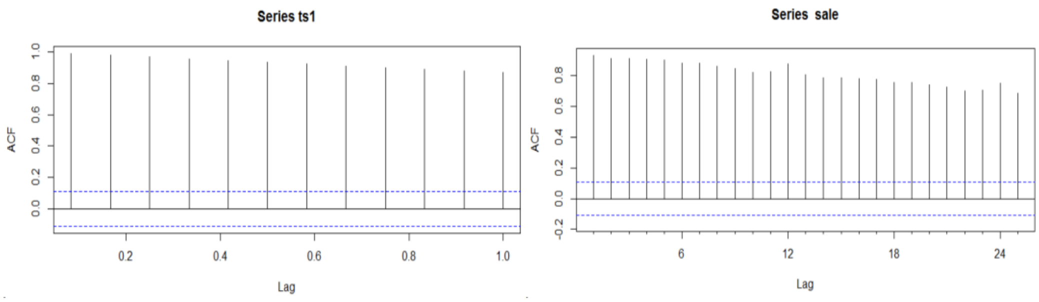


Figure 3



It was clear from the ACF plot as well as the time plot for the adjusted and non-adjusted data that this data is not stationary. The box-Ljung test showed this as well, as we were able to reject the null hypothesis that the correlation are equal to zero up to lag K for both the adjusted and non-adjusted dataset. Because of this we looked to several data transformations. The data transformation that had the best results in terms of stationary was the log returns. The ACF plots for the adjusted and non-adjusted data are below in figure 4.

Figure 4



We conducted multiple tests to see if the log return data was stationary are not. From the Jarque-Bera test, it was seen the p-value was lower than the alpha 0.05 for both the adjusted and non-adjusted dataset, so we can go ahead and accept the null hypothesis that the distribution is normal. Along with that, the Ljung box test accepted the null hypothesis that the autocorrelations up to lag k are equal to zero, since the p-value was less than the alpha 0.05. Because of this as well as the ACF plots showing autocorrelations close to zero other than a few occasions in the non-adjusted data we can say this data was stationary. It was clear that the adjusted dataset was more stationary and had more visible trends, thus was the main dataset used in this study. The test and ACF plots are below.

Figure 5

Jarque-Bera normality test

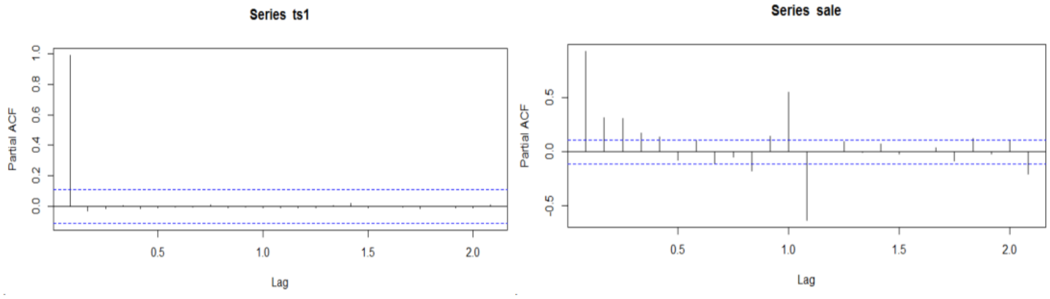
	Adjusted	Non Adjusted
p-value	0.0008011	0.00472
Result	Reject H0 null hypothesis of normal distribution because p-value is <0.05	

Figure 6

Ljung box test for independence of white noise

	Adjusted	Non Adjusted
p-value	0.00000000000000022	0.00000000000000022
Result	Reject H0 of zero correlation at 5% significance level because p-value < 0.05	

Figure 7



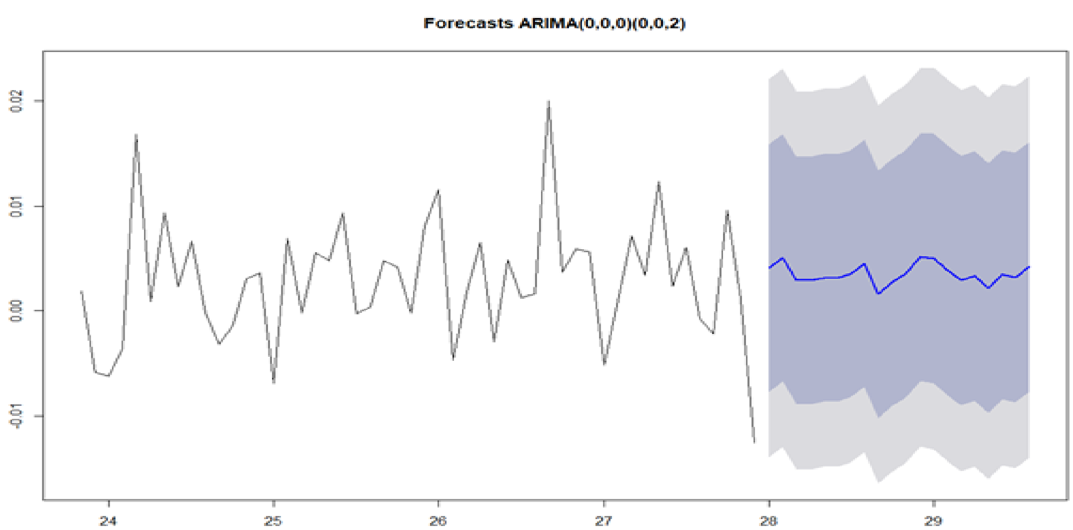
MODELS

Model 1:

For the first model, we utilized the Auto ARIMA function using the BIC criterion to find the most optimal ARIMA model. The adjusted retail sales time plot showed a non-stationary time series, so to convert it to a stationary time series, we used the log returns time plot instead. As stated above, the log returns was a stationary time series that would be better to use for forecasting.

From the Auto ARIMA function, it specified that the best model based on the BIC criterion is the SARIMA0,0,0)(0,0,2). This model does not contain any autoregressive or moving average components, as the first set of values are all set to zero. This model only contains seasonality effects. The plot of the forecast is below.

Figure 8



In the forecast plot, it can be seen that the increase and decrease in the log returns is from the seasonality effect of the SARIMA model. As the seasonality was low from this data, it can be seen the increase and decreases in value follow no consistent pattern.

Based on the ACF plot of the residuals, it was seen that the ACF decays to zero and is consistently near zero for all of the lags. Along with the ACF plot the Box-Ljung test for the residuals accept the null hypothesis that the residuals are not a lack of fit, while the histogram shows the residuals are normally distributed. Based on all of these factors, we can say this is an adequate model due to the residuals check.

Figure 9

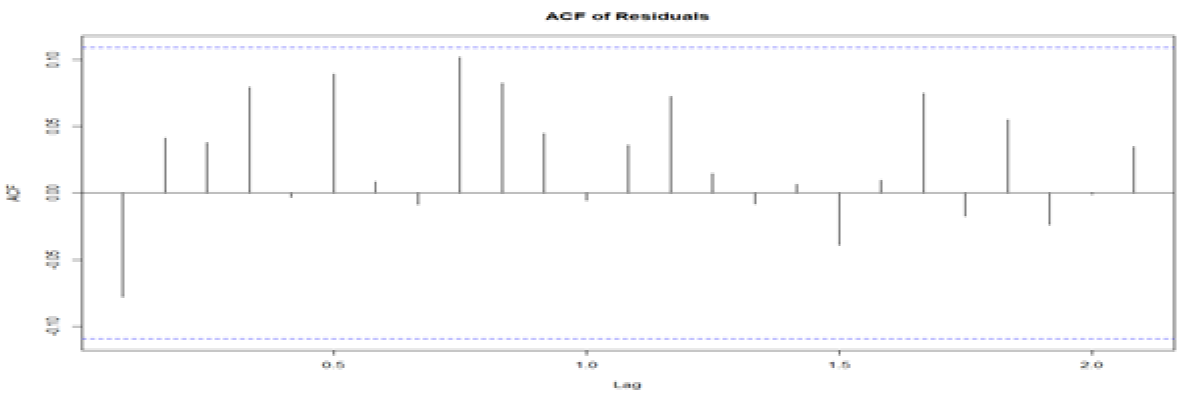
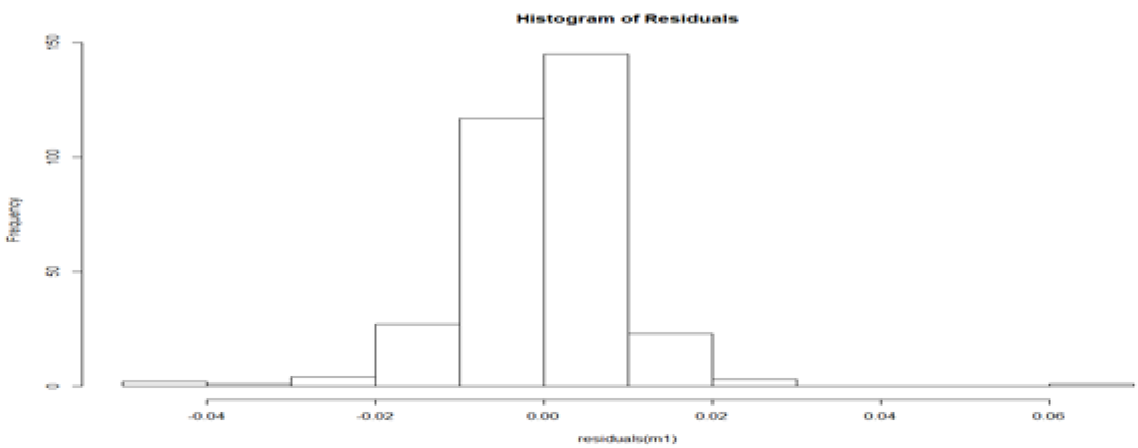


Figure 10



The results from the backtest is below. We can see from these performance metrics that the model was good as the RMSE and mean absolute error were very low. Along with that, the mean absolute percentage error was fairly low at 2.190401%. Based on the residuals check and the performance metrics, we can consider this time series analysis as an accurate and adequate forecast.

Figure 11

RMSE	Mean absolute error	Mean Absolute Percentage error
0.01022818	0.007300325	2.190401%

We checked to see what the effect of the unadjusted dataset would have on the model chosen from the Auto-Arima function. An SARIMA(2,0,0)(0,1,4) model was chosen and it performed well with the residual checks. But it did not perform as well as the SARIMA(0,0,0)(0,0,2) model for the adjusted data. Which is a reasonable conclusion since the unadjusted dataset is prone to more volatility and is not as smooth as the adjusted dataset.

Model 2:

The second model looked at are through the EACF plot. We are still looking for an ARIMA model and will use both the ACF/PACF plots, as well as the EACF plot to determine what ARIMA or SARIMA model would be most efficient. Below are the ACF/PACF and EACF plots.

Figure 12

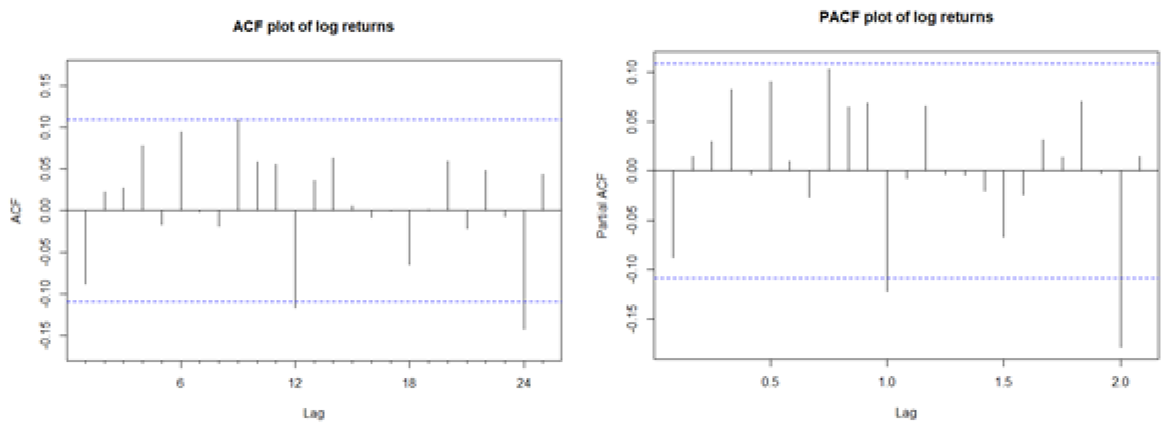


Figure 13

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	o	o	o	o	o	o	o	o	o	o	o	x	o	o
1	x	o	o	o	o	o	o	o	o	o	o	o	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	o	o
3	x	o	x	o	o	o	o	o	o	o	o	o	o	o
4	o	x	x	o	o	o	o	o	o	o	o	o	o	o
5	o	x	x	x	o	o	o	o	o	o	o	x	o	o
6	o	x	x	x	x	o	o	o	o	o	o	o	o	o
7	x	o	x	x	x	o	o	o	o	o	o	o	o	o

ACF and PACF plots of log returns show significant negative correlations at lag 12 and 24, they indicate annual seasonality but very weak as they are just around range -0.1 and -0.15.

Dickey-Fuller test with p-value = 0.01 and Ljung box tests at lag 6 and 12 have p-value 0.24 and 0.09 respectively show the series already stationary.

So according to the EACF triangle, I build a model with the AR,MA parts to be (0,0,0) but in order to account for the seasonality shown in ACF and PACF plots (at lags 12 and 24), I include some seasonal components in the model, hence the seasonal part is (1,0,1) with period = 12 representing annual seasonality.

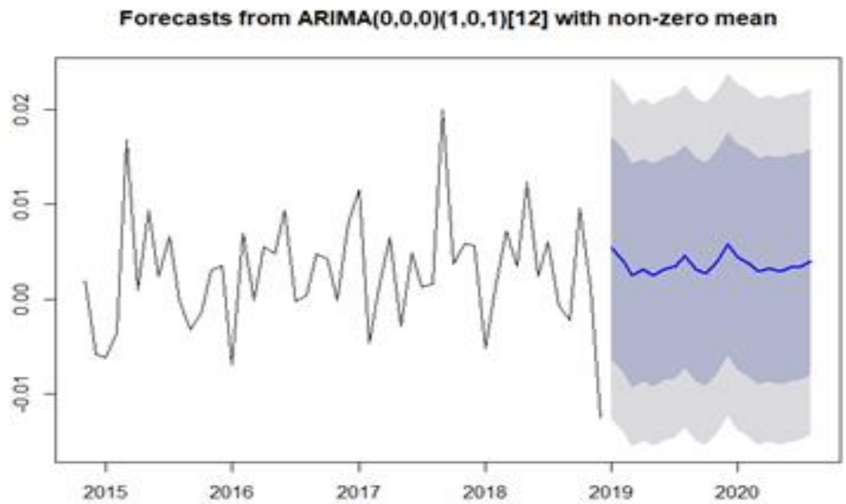
Figure 14

```
ARIMA(0,0,0)(1,0,1)[12] with non-zero mean

Coefficients:
        sar1     sma1    mean
      0.5304  -0.6897  0.0034
s.e.  0.2133   0.1855  0.0004

sigma^2 estimated as 0.0000849:  log likelihood=1056.71
AIC=-2105.42   AICc=-2105.29   BIC=-2090.31
```

```
z test of coefficients:

            Estimate   Std. Error z value  Pr(>|z|)
sar1       0.53039512  0.21334612  2.4861 0.0129160 *
sma1      -0.68974210  0.18547541 -3.7188 0.0002002 ***
intercept  0.00344021  0.00035776  9.6159 < 2.2e-16 ***
```

The SARIMA(0,0,0)(1,0,1)[12] model has log-likelihood of 1,056.71 and AIC of -2,105.42 and all coefficients to be significant.

Figure 15



The forecast went down deeper at first but after that goes same trend as Auto arima model, but with smoother movements.

Backtest results show RMSE is around 0.0055, not very small, considering the standard deviation of log returns is 0.9% and the data moves in range from -3.9% to +6.5% (min to max log returns). However, Mean absolute percentage error is good as it is only 0.32%.

Figure 16

RMSE	Mean absolute error	Mean Absolute Percentage error	Symmetric Mean Absolute Percentage error
0.005460706	0.004000081	0.3194331	0.9474797

The residuals look good), shown by ACF plot and tested by Ljung box test at lag 6 and 12 to have no autocorrelations, are white noise series, indicating a well fit model.

Because exploratory analysis showed that Retail Sales series is non stationary the GARCH analysis starts with applying log transformation to Retail Sales time series and calculation of returns.

Figure 17

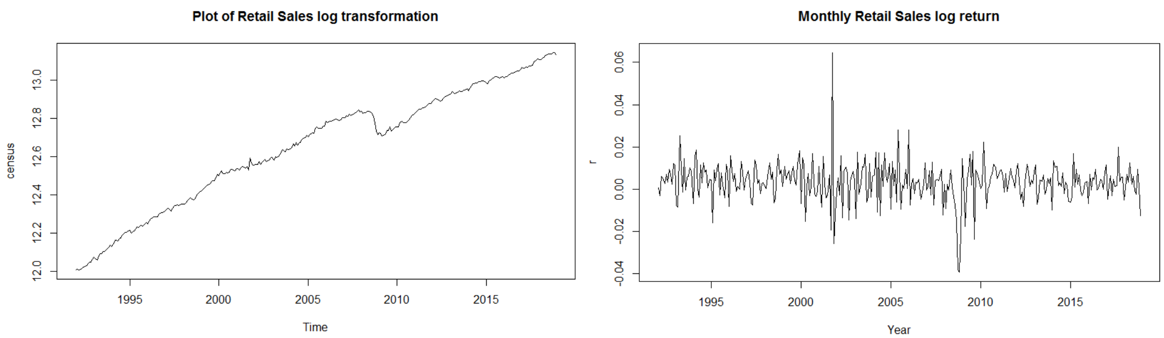
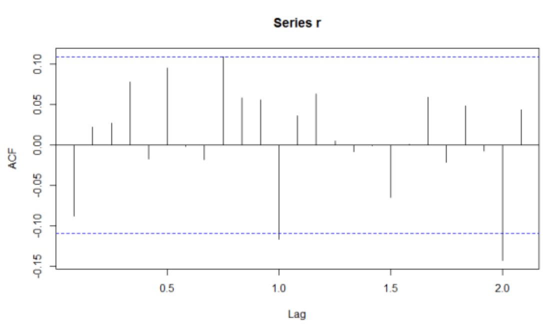


Figure 18

	Adjusted
skewness	0.0698
kurtosis	7.5570
Jack Berra normality test	0.00000000000000022
Results	Not normal

Above table show that Adjusted Retail Sales log returns are not normal. Autocorrelation plot of log returns and p-value equal to 0.15 of Ljung test show that log returns are not correlated. Analysis of EACF plot log returns tell us that possible model could be ARIMA(0,1,0). Auto Arima best model ARIMA(0,1,0)(0,0,2)[12] with drift

Figure 19



AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	o	o	o	o	o	o	o	o	o	o	o	x	o	o
1	x	o	o	o	o	o	o	o	o	o	o	o	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	o	o
3	x	o	x	o	o	o	o	o	o	o	o	o	o	o
4	o	x	x	o	o	o	o	o	o	o	o	o	o	o
5	o	x	x	x	o	o	o	o	o	o	o	x	o	o
6	o	x	x	x	x	o	o	o	o	o	o	o	o	o
7	x	o	x	x	x	o	o	o	o	o	o	o	o	o

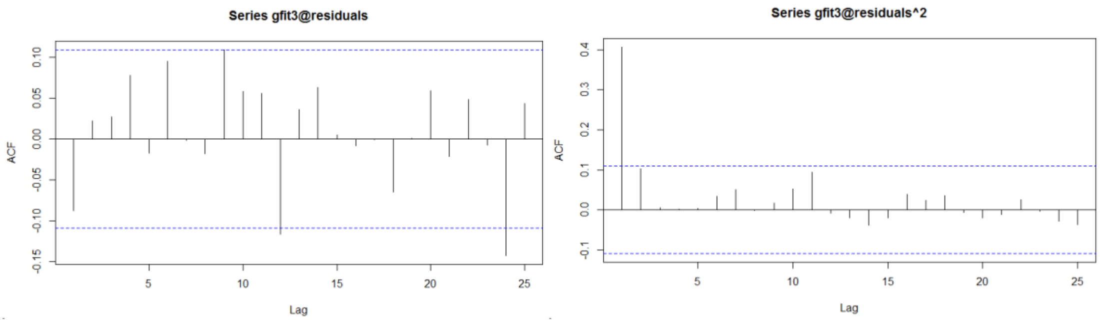
Fitting this model shows that only drift is significant. Because there is a autocorrelation at lag 1 on squared log returns the model is adequate for further GARCH analysis.

Model 3:

First we apply GARCH(1,1) on top of ARMA(0,0). ACF shows that residuals are not correlated, but squared residuals are correlated, this is confirmed by Ljung-Box test (p-value 0.1515 and

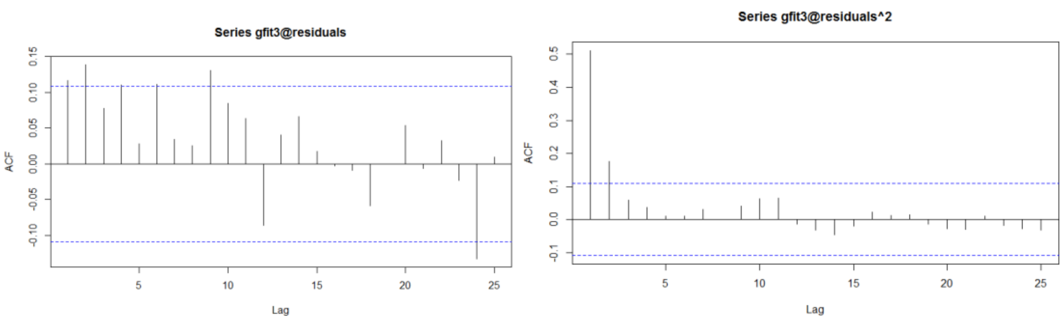
0.00000006392), Jarque Berra test shows that both residuals and squared residuals are not normal.

Figure 20



Second, we apply GARCH(1,1) on top of ARMA(0,2) returned by auto arima, ACF shows that residuals are correlated, and squared residuals are also correlated, this is confirmed by Ljung-Box test (p-value 0.0.00000006392, 0.000000000000004774), Jarque Berra test shows that both residuals are not normal.

Figure 21



Predictions for GARCH (1, 1) + ARMA (0, 0) on the left, and for GARCH (1, 1) + ARMA (0, 2) on the right

Figure 22

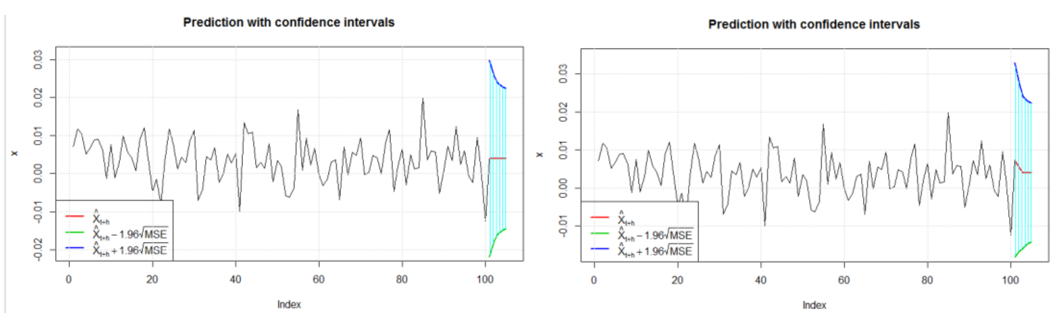


Figure 23

Backtest results

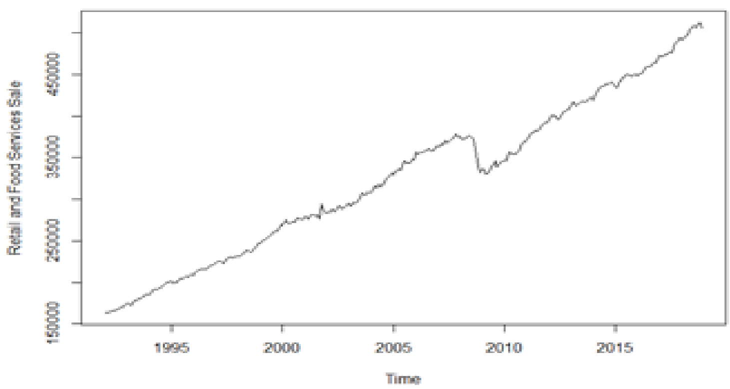
	RMSE	Mean absolute error	Mean Absolute Percentage error	Symmetric Mean Absolute Percentage error
GARCH(1,1) + ARMA(0,0)	0.0073	0.0056	2.1363	1.2620
GARCH(1,1) + ARMA(0,2)	0.0073	0.0053	1.9823	1.1723

Both GARCH models did not provide accurate volatility prediction, because squared residuals were correlated, this suggest that there must a better model, which is not captured by above analysis.

Intervention Model:

From the time plot, clearly there is a significant drop in retail and food services sale estimated to be around July 2008, which was caused by the 2008 recession.

Figure 24



ACF and PACF plots of Retail and Food Services Sale show non-stationary and AR(1) model. However, Dickey-Fuller test has p-value of 0.99, so we cannot reject the null hypothesis of unit root , together with EACF plot suggests a differencing of the series and an ARIMA(0,1,0) model. Besides, ACF and PACF plots of the differencing of the sale both show spikes at lag 12 and 24. So we include seasonal components AR(1), MA(1) and period 12 in the model.
The SARIMA(0,1,0)(1,0,1)[12] model has log likelihood = -3,056.75 and AIC = 6,119.49. Residuals are tested to be stationary. Now, we hypothesize a permanent change in level due to the 2008 recession by using a step function and no lags. Compared to the original model, the new model has lower log likelihood = -3,061.36, higher AIC = 6,128.72. Residuals look good and are tested to be white noise.

Figure 25

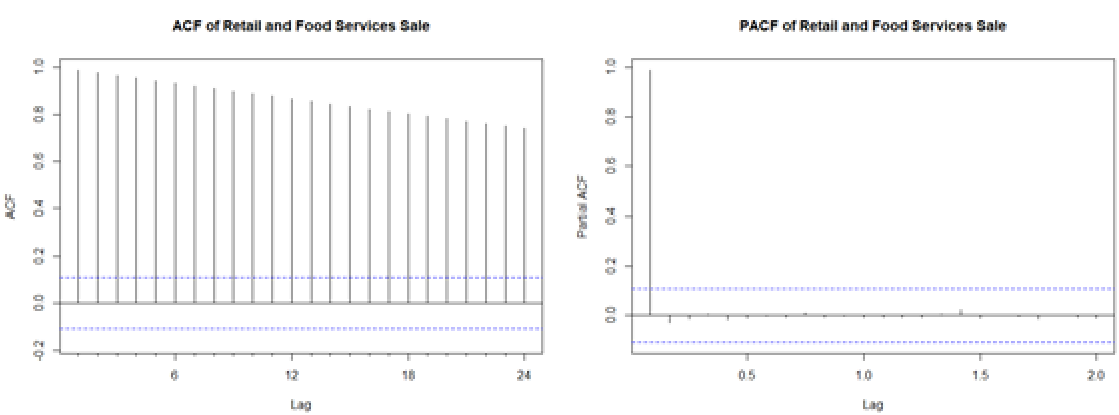
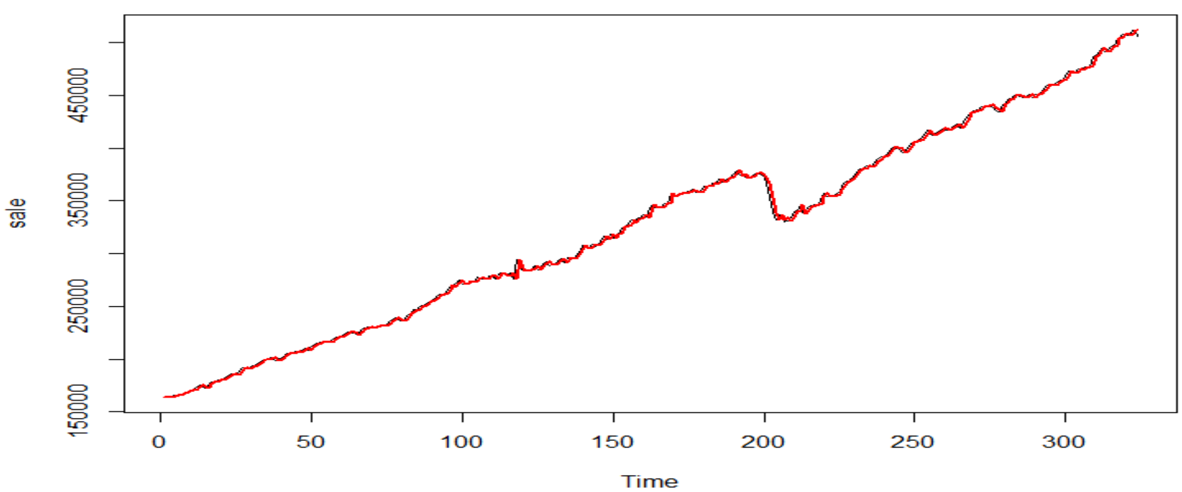


Figure 26

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	o	o	o	o	o	o	o	o	o	o	o	o	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	o	o
3	x	x	o	o	o	o	o	o	o	o	o	x	o	o
4	x	x	x	o	o	o	o	o	o	o	o	x	o	o
5	x	x	x	o	o	o	o	o	o	o	o	x	o	o
6	x	x	x	o	o	o	o	o	o	o	o	o	o	o
7	x	o	x	x	o	o	o	o	o	o	o	o	o	o

Then, we hypothesize a temporary change in level due to the 2008 recession by using a step function and no lags. Compared to the permanent model, this new model has lower log likelihood = -3,061.79, higher AIC = 6,129.58. Residuals still look good and are tested to be white noise. Looking closely at the drop point, it can be seen that the temporary model does not fit as well as the permanent model, might be there is not enough time to see the return to the pre-endorsement behavior.

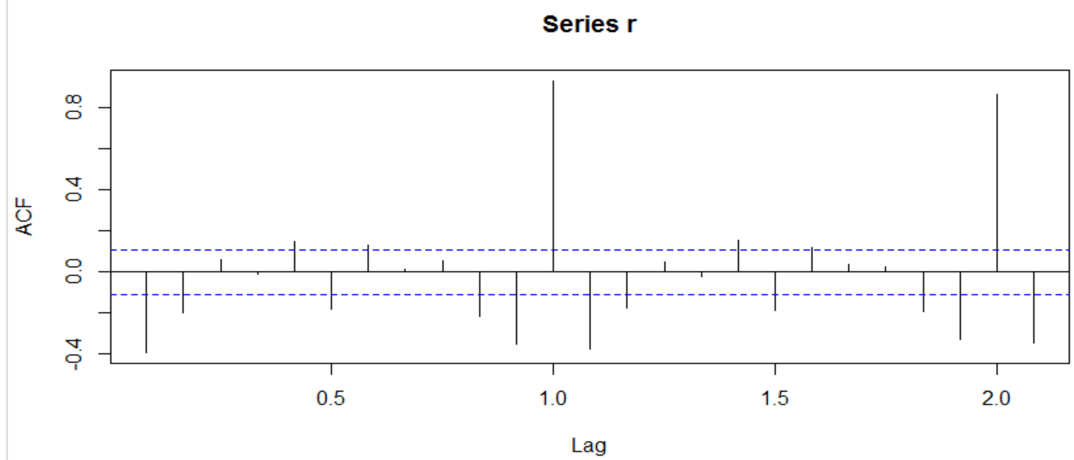
Figure 27



Non-Adjusted Model:

Here are results for a model run on non-adjusted data, there was a seasonal effect discovered in acf plot of returns. This model was considered to be the best out of not only the unadjusted dataset, but also of all the models from the adjusted dataset. The clear seasonality can be seen at lag 1 and lag 2, as there is a small negative ACF value the lag before, followed by a large value at lag 1 and lag 2, and followed by another small negative value after.

Figure 28



The results of the test are below. It was seen that the residuals passed all of the test, while the mean absolute percentage error was very low at 0.0028%. This shows this model does a good job in catching the trends and increasing linear path the time series follows. From the time series forecast plot in figure (), it was seen that it follows a constant increasing pattern that does a good job in predicting the log return.

Figure 29

ARIMA(0,1,13)	score	result
Jarque Bera Test	0.2126	normal
Box-Ljung test	0.00000000000000022	#reject hypothesis of zero correlation, there is correlation
Box-Ljung test residuals	0.1909	Residuals are not correlated

Figure 30

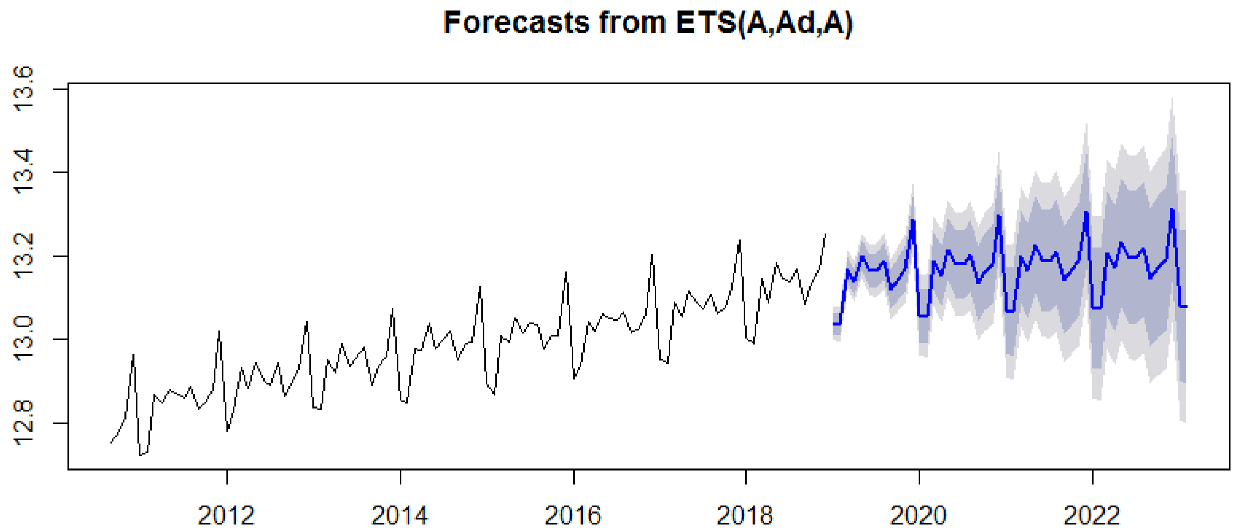


Figure 31

Backtest

	RMSE	Mean absolute error	Mean Absolute Percentage error	Symmetric Mean Absolute Percentage error
ARIMA(0,1,13)	0.0461	0.0372	0.0028	0.0028

Interpretation Best Result

With all of the models we saw good results, as the RMSE and MAE were low, which leads to a lower mean absolute percentage error. Out of all of the models, the ARIMA(0,1,13) had the lowest mean absolute percentage error overall between both the adjusted and non-adjusted dataset, as this model was trained from the non-adjusted dataset. All of the other models were trained from the adjusted dataset and we can see that the SARIMA(0,1,0)(1,0,1) had the lowest mean absolute percentage error. The table for the comparison of the performance metrics are below.

Figure 32

Comparison of models

	RMSE	Mean absolute error	Mean Absolute Percentage error	Symmetric Mean Absolute Percentage error
GARCH(1,1) + ARMA(0,0)	0.0073	0.0056	2.1363	1.2620
GARCH(1,1) + ARMA(0,2)	0.0073	0.0053	1.9823	1.1723
ARIMA(0,1,13)	0.0461	0.0372	0.0028	0.0028
SARIME(0,0,0)(0,0,2)	0.01022818	0.007300325	2.190401	
SARIMA(0,1,0)(1,0,1)[12]	0.005460706	0.004000081	0.3194331	0.9474797

Overall, we can say that all of the models performed adequately for the adjusted data and would be viable ways to predict the values for the log returns in the adjusted data. With that being said the SARIMA (0,1,0)(1,0,1), performed the best and will minimize error the most optimal way in the forecasts. The

Conclusion:

It can be concluded that the SARIMA(0,1,0)(1,0,1)[12], performed the best and was significantly better than the other models for the adjusted dataset. For the unadjusted dataset the ARIMA(0,1,13) performed very well, and had a lower mean absolute percentage error than the SARIMA(0,1,0)(1,0,1)[12]. This means that the SARIMA model for the unadjusted dataset should be used on the adjusted dataset to see if it has similar performance metrics. The intervention model did well in putting in the recession of 2008 into consideration, but there is not enough information with it to say it is a better model than the models tested in this study. The intervention model should be looked into more detail to see if it would be possible to create an intervention model that has better mean absolute percentage error than the models used throughout this study.

Non-technical Summary

This project examines a time series of 324 monthly sale observations of Monthly Sales for Retail Trade and Food Services in the U.S from Jan 1992 to Dec 2018 in an attempt to build a model capable of predicting the future monthly sales.

The data was provided by the United States Census Bureau, which is a principal agency of the U.S. Federal Statistical System, responsible for producing data about the American people and economy. The data was processed to represent monthly sales of three million retail and food services firms.

Sales are highly affected by seasonal aspects, for example, holidays' shopping and changes between hot/cold weather can boost or shrink sales of specific firms (retailers selling holiday decorations, warm clothes, summer clothes...). These seasonal influences in this particular data tend to occur repeatedly every 12 months and have been measured by U.S Census Bureau. Hence, the agency provided two versions of data, one is monthly sales with seasonal effects removed (called adjusted sales), and the other is original monthly sales with seasonal effects incorporated (called non-adjusted sales).

We built prediction models for both time series. To prepare a stable data for models, we had to transform the sales so that each monthly sale follow each other on a smooth movement throughout the years, besides, we based the models on the percentage changes between two consecutive months.

For the adjusted sales series, we came up with a model that could predict the percentage change in future monthly sales with 99.68% accuracy and variations of ±0.5%.

For the non-adjusted sales series, the model could predict the percentage change in future monthly sales with 99.99% accuracy and variations of ±4.6%.

Also, we observed a clear drop in retail and food services sales after the big recession of 2008, which we estimate to have occurred for this data set in July 2008. Thus, we did an analysis on the drop and built another model to incorporate the permanent effect of the 2008 recession onto this time series.

Appendix:
Dominick Amalraj's Individual Paper:

For this project, I have contributed heavily in the exploratory analysis, the auto-arima model formed, and the overall interpretation of the results. For the exploratory analysis, I conducted analysis on the distribution, normality, was the data seasonal and if the data was stationary. I looked into the exploratory analysis for both the adjusted and non-adjusted dataset, as well as dove into the raw data, log of the data, and log returns for both the adjusted and non-adjusted datasets. From going in depth into both the adjusted and non-adjusted datasets, our team was able to get more direction on what type of data to use and the data transformation that makes the most sense. The time plot of the data, ACF/PACF plots, and histograms of the data were the main visual analysis I created in the exploratory analysis. The tests that I ran include the dickey-fuller test, Jarque-Bera test, and the Ljung-box test.

The model that I worked on was the SARIMA model from the auto-arima function with the BIC criterion. After finalizing that the log returns of the adjusted data will be used I utilized the auto-arima function with this data to come up with my forecasting model. Along with creating the model, I had an in-depth residual check to make sure this was an adequate model. For the residual check I looked into if the residuals were normally distributed, were there any autocorrelation, were the residuals stationary, and if the residuals were independent of each other. I utilized the ACF plots to see if the residuals were highly correlated with previous lags, and the histogram to visualize the distribution of the residuals. The Ljung-box test was used to determine if the residuals were stationary and the Jarque-Bera test and Dickey-Fuller test was used to determine the skewness and normality of the residuals.

Along with my contributions in the analysis, I gave the whole presentation of all of the information including the work of Long and Michal. This was a challenging part of the project, because not only did I had to understand my applications well, but I had to understand the application Long and Michal created to be able to tie everything together in an effective manner. I had to get a good understanding of the work they did and understood why they did certain actions to better be able to tell the story of our process.

Lastly my other major contribution was the interpretation of the results and the conclusion. This was necessary so we can wrap the whole project up and get the main insights. To be able to do this, again being able to understand fully the process my partners toke was vital to be able to make insightful conclusions. The conclusion was needed to close the project and leave the audience with the final takeaway of the project.

This was an interesting project to work on. One of the main concepts I learned was the use of data transformation is vital in creating effective forecasts with non-stationary data. Our raw data for both the adjusted and non-adjusted dataset were not time-series friendly as it was a linear increasing trend. By taking the log returns we were able to make it more stationary and be able to create forecasts that were accurate and impactful in predicting the values. Along with the work I have done, I was present and contributed in every group meeting to make sure all members were on the same page.

Michal Chowaniak

DSC 425 Individual Report

Individual Report (40 pts)

Each student should submit an individual report on their contribution to the team's efforts. You should detail the duties that you completed and the analyses that you either conducted yourself or contributed to. The report should be at least a full page long with detailed descriptions of your contribution. (20 pts)

My contribution to the group project:

1. Research and proposed alternative time series Amazon stock, but we choose Retail Sales series.
2. Preliminary analysis group submission: Time series plot, Trends, Time segments, Seasonality, Decomposing function,
3. Hang out video call, discussion 02/14/19- choosing time series
4. Preliminary analysis: each member worked on his own analysis

    Time series, plots, trends, time segments, seasonality, decomposing, histograms, normality tests, independence test, acf, pacf, stationarity, Dickey-Fuller test, auto Arima ARIMA(0,1,0)(0,0,2)[12] with drift, arima (1,1,1), residuals, predictions, backtesing.
5. Hang out video call – 03/02/19 – discuss plan for next steps for project
6. Full GARCH model on adjusted:
    a) log transformations, normality, acf, pacf, stationarity test, acf returns, eacf, ARIMA(0,1,0), coeficietns, residual analysis, box test, forecast bactest
    b) auto arima, residuals, backtest, acf on r^2, (abs)r
    c) standard GARCH, acf on r^2, (abs)r, GARCHJ(0,1), residuals, GARCH(1,1), residuals, normality, independence, acf,
    d) fGARCH(1,1) residuals, ARMA(1,1)+GARCH(1,1), residuals, predictions
7. Full GARCH model on non adjusted data
    a) log transformations, normality, acf, pacf, stationarity test, acf returns, eacf, ARIMA(0,1,1), coeficietns, residual analysis, box test, forecast back test
    b) auto arima, ARIMA(0,1,0)(0,0,2)[12] with drift, residuals, back test, acf on r^2, (abs)r
    c) standard GARCH, acf on r^2, (abs)r, GARCHJ(0,1), residuals, GARCH(1,1), residuals, normality, independence, acf,
    d) fGARCH(1,1) residuals, ARMA(1,1)+GARCH(1,1), residuals, predictions

8. Hang out video call – 03/07/19 -plan for project and presentation
9. Presentation - GARCH model – 4 slides
10. Presentation - comparison of 3 models slides – 2 slides
11. Hang out video call – 03/14/19 – plan for final report
12. Final Report - Introduction – interesting facts about Retail Sales and Food Services data series from United Census Bureau.
13. Final Report - Exploratory Analysis

a) Plots of adjusted and non adjusted, histograms, decompose, Jarque-Bera normality test, Ljung box test, stationarity, autocorrelation

14. Another GARCH analysis
    a) log transformations, normality, acf, pacf, stationarity test, acf returns, eacf, ARIMA(0,1,1), coefficients, residual analysis, box test, forecast back test
    b) auto arima, ARIMA(0,1,0)(0,0,2)[12] with drift, residuals, back test, acf on r^2, (abs)r
    c) fGARCH(1,1) residuals, ARMA(0,2)+GARCH(1,1), residuals, predictions, GARCH backtest
15. Final Report - GARCH Analysis based on GARCH #14
16. For Final report - Full model on non adjusted data as a comparison to our adjusted data models
    a) Log time series, differences, acf, seasonality, differences x 12, acf, arima(r, order=c(0,0,13), fixed=c(0,NA,NA,0,0,0,0,0,0,0,NA,NA,NA,NA)), stationarity, normality, correlation, residuals, acf, box test, predictions, back test
    b) arima(census, order=c(0,1,13)), coefficients, returns stationarity, returns normality, residuals, acf, box test, predictions, forecast, back test

A paragraph of takeaways that you have learned from the project. The paragraph should be detailed and specific about something positive that you have learned about time series analysis or data analysis in general. (5 pts)

Time series analysis was a new subject to me. It is very challenging and I feel I only know a little about it. The most important thing I learnt from the project, except the usual things like checking for normality, seasonality, analyzing autocorrelation, working on residual analysis, finding out type of a model, doing forecast and back testing, was to not trust the dataset. Our group spent most of time trying to extract a proper model from our seasonally adjusted data set, and at the end it turned out that we were looking at a wrong place, because my model build for comparison on non-adjusted dataset had an error rate more then hundred times better than our best model built on seasonally adjusted data set. Also the best way to learn time series for me was to work on my own redoing the same problem over and over.

Individual report

Long Bao Nguyen

The duties i either contributed or conducted myself to:

Group forming:

- Posted on D2L Discussion to form group.

Dataset:

- Suggested the Monthly Sales for Retail Trade and Food Services dataset.

Exploratory analysis: on both adjusted and non-adjusted data

- Time plots, histograms, qq-plots.
- Distribution tests (Skewness, Kurtosis, Jarque-Bera).
- Decompositions.
- ACF, PACF, Ljung-Box tests of white noise.
- Seasonality difference.

Presentation: on EACF model slides

- Model fitting.
- Residuals analysis.
- Forecast and backtests.

Group Project Report:

- EACF - SARIMA(0,0,0)(1,0,1)[12] model write-up on model fitting, residuals analysis, forecast and backtests.
- Intervention analysis write-up on model fitting, residual analysis of SARIMA(0,1,0)(1,0,1)[12] model, permanent and temporary intervention models.

Non-technical report:

- Write-up.

Appendix:

- Graphs of SARIMA(0,1,0)(1,0,1)[12] model, permanent and temporary intervention models residuals.

Takeaway

Long Bao Nguyen

The project gave me chance to apply the new concepts about different types of time series (random walk, white noise...), stationary, seasonality, dependence and intervention effects to real data and interpret them. Log-transformation, differencing and volatility are not new to me but seeing how they could highly affect the performance of models, i could understand when to apply them better now, learning about ARCH/GARCH model also proves useful in volatility modeling. Besides, i got to practice coding many types of plots and tests to assess time series and base on those results (reading ACF,PACF,EACF plots and tests results) to build AR, MA, ARMA, ARIMA, SARIMA models for the purpose of forecasting and which model is specifically fitted for the data that i am working on.