# 5550 Project

Ryan Ulring

4/22/2021

i. Introduction

The data I will be using for this analysis contains the U.S. city average consumer price index (CPI) for all urban consumers, excluding food and energy, from January 2011 to February 2021. I found this data from the U.S. Bureau of Labor Statistics, with a series id of "CUUR0000SA0L1E" and a series title of "All items less food and energy in U.S. city average, all urban consumers, not seasonally adjusted." The goal of this analysis is to find an appropriate model that fits the data well.

In the first plot below, there is a clear upward trend in the average consumer price index over time. Based on this plot, the trend also appears to be fairly linear. There is also a seasonal trend in the data, which is hard to picture, so I plotted the differenced data in the second plot below in order to show the seasonal trend. In the differenced plot, we can more clearly see the seasonal trend that appears throughout the data, and it does not seem to have a specific parametric form. Since there appears to be a linear trend in the data, and since the seasonal component does not appear to have a specific parametric form, I believe that a seasonal means model with a linear trend is an appropriate model here.

**U.S. City Avg CPI for All Urban Consumers, Excluding Food & Energy**
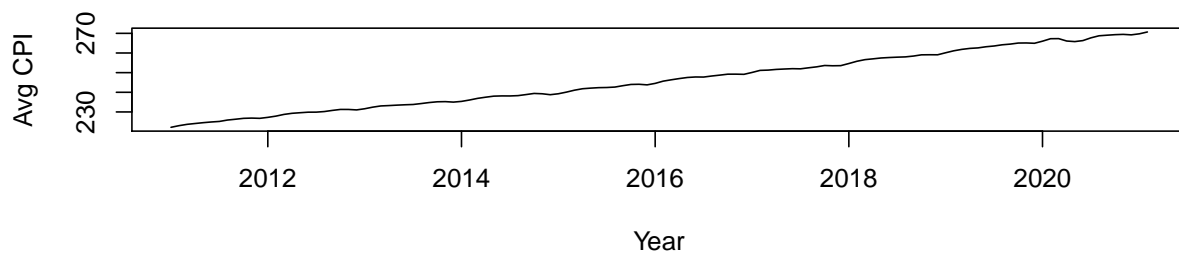

Differenced Data
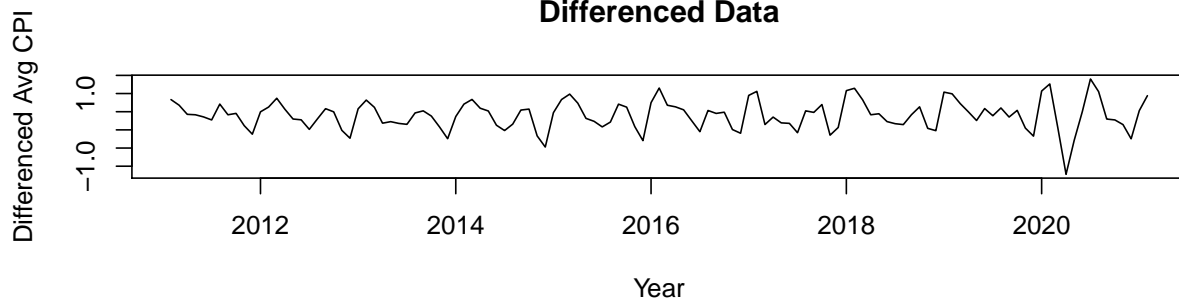
ii. Models for Trend and Seasonality

For this next part of my analysis I will explicitly estimate the trend and seasonal components, detrend and deasonalize the data, and use ARMA modeling techniques to analyze the detrended, deseasonalized data.

First, I will look at the detrended, deseasonalized data assuming a seasonal means model with a linear trend, in order to make sure this model removes the trend from the data.
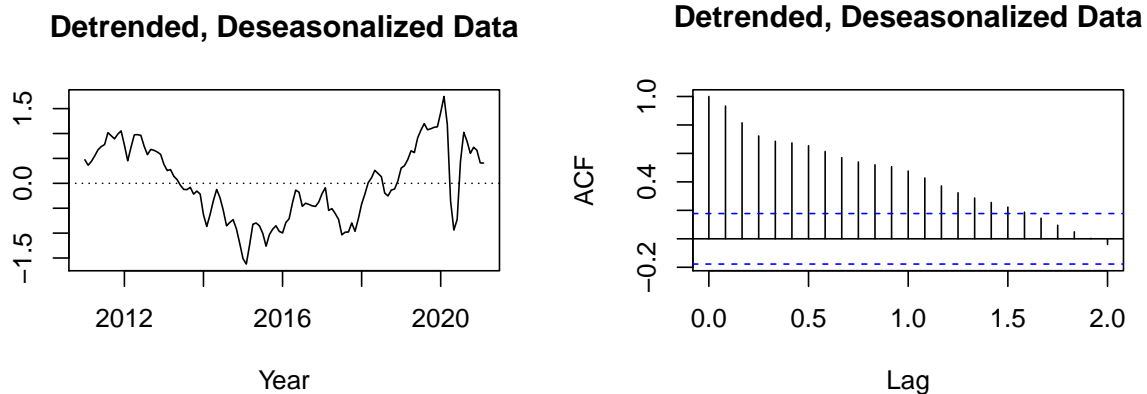

For this series, there does appear to be a noticeable pattern(mean does not appear to be constant), so this model is probably not appropriate. Also, the sample acf plot does not decay rapidly, which is further evidence that this model is not appropriate. Since this detrended, deseasonalized data has somewhat of a U-shaped pattern, I will next look at the detrended, deseasonalized data assuming a seasonal means model with a quadratic trend.
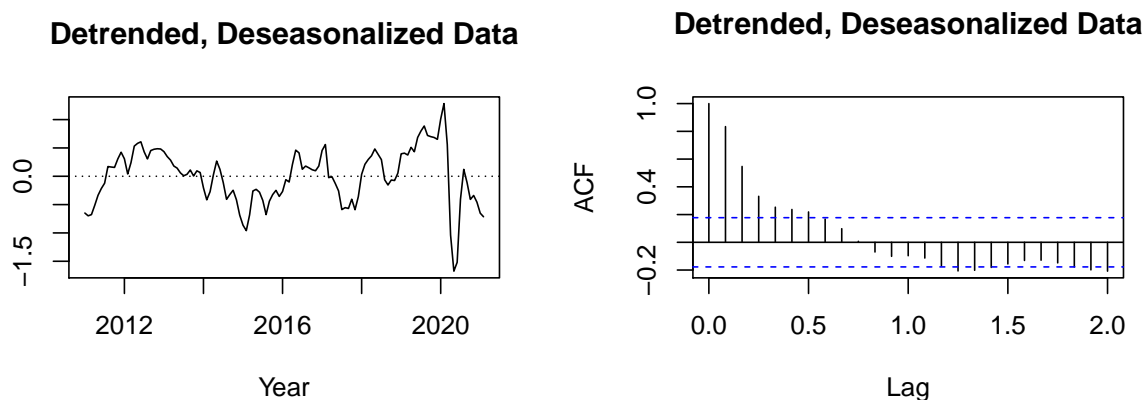

Here, we can see that there is now no noticeable pattern in the detrended, deseasonalized data(mean and variance appear to be fairly constant). In addition to this, the sample acf plot seems to decay fairly rapidly, so this may be a good model to use. Since it appears that this model has successfully removed the trend in the data, and the detrended, deseasonalized data appears to be relatively stationary, so I will continue with this model.

Next, I will look at ARMA models to analyze the detrended, deseasonalized data. In order to get an idea of what ARMA models I should consider, I will look at the sample acf and pacf plots of the detrended, deseasonalized data.

**Detrended, Deseasonalized Data**   **Detrended, Deseasonalized Data**
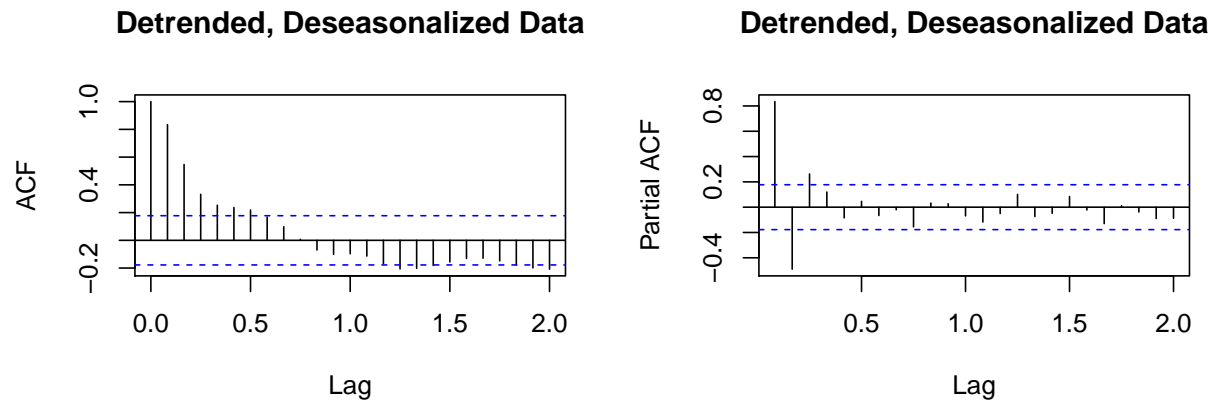
From the sample acf plot, it appears that the acf tails off. The sample pacf plot shows us that the pacf appears to either be cut off after the first few lags, which indicates an AR model, so I will try to fit AR(2), AR(3), and AR(4) models to the detrended, deseasonalized data. First I will look at the fit for the AR(2) model.
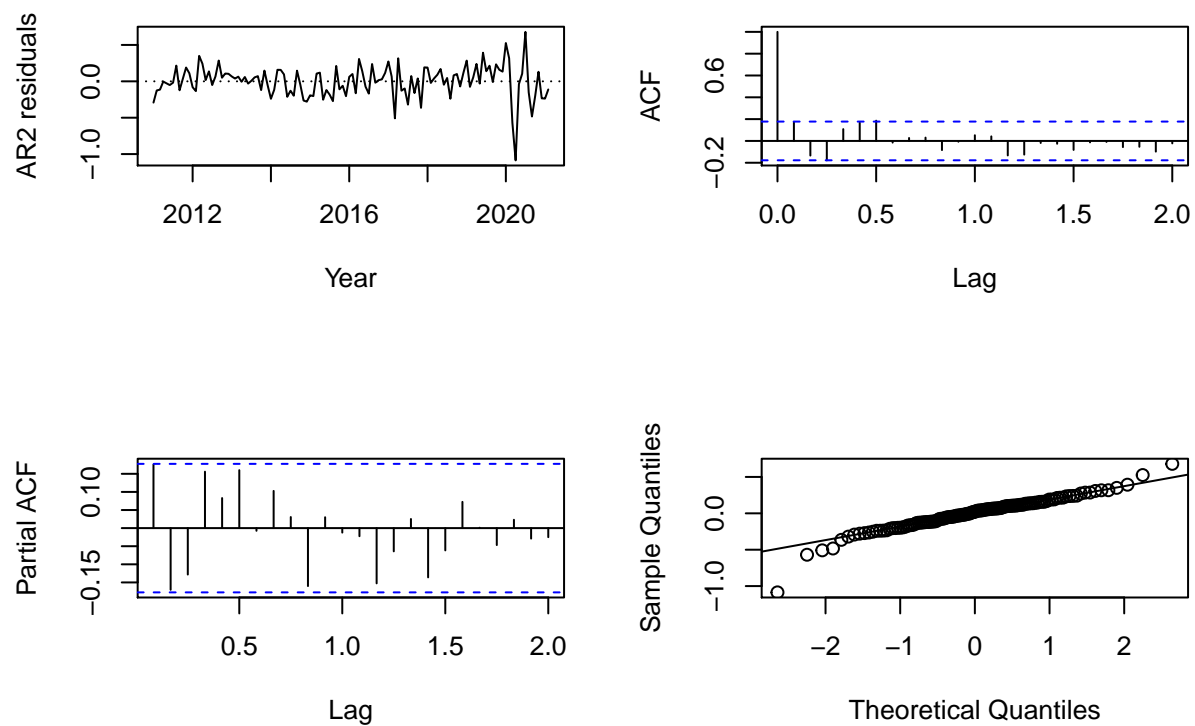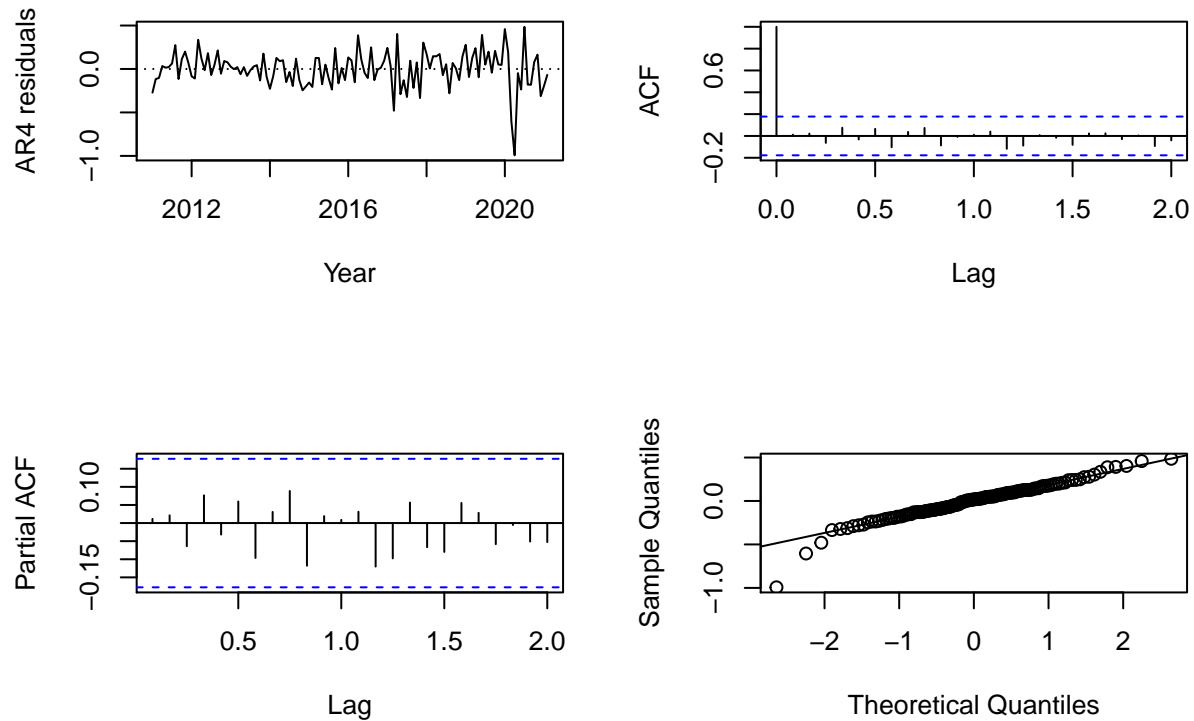
```
##
##  Box-Ljung test
##
## data:  ar2.model$residuals
## X-squared = 29.960979533123663, df = 24, p-value = 0.18604844753943106
```

3

For this model, the residuals do appear like they could be white noise, the acf and pacf plots do not have many lags outside of the significance threshold, and the QQ-plot seems to fit well; however, the p-value for the Ljung-Box test is small, which indicates that this series is not white noise. Next I will look at the fit for the AR(3) model.
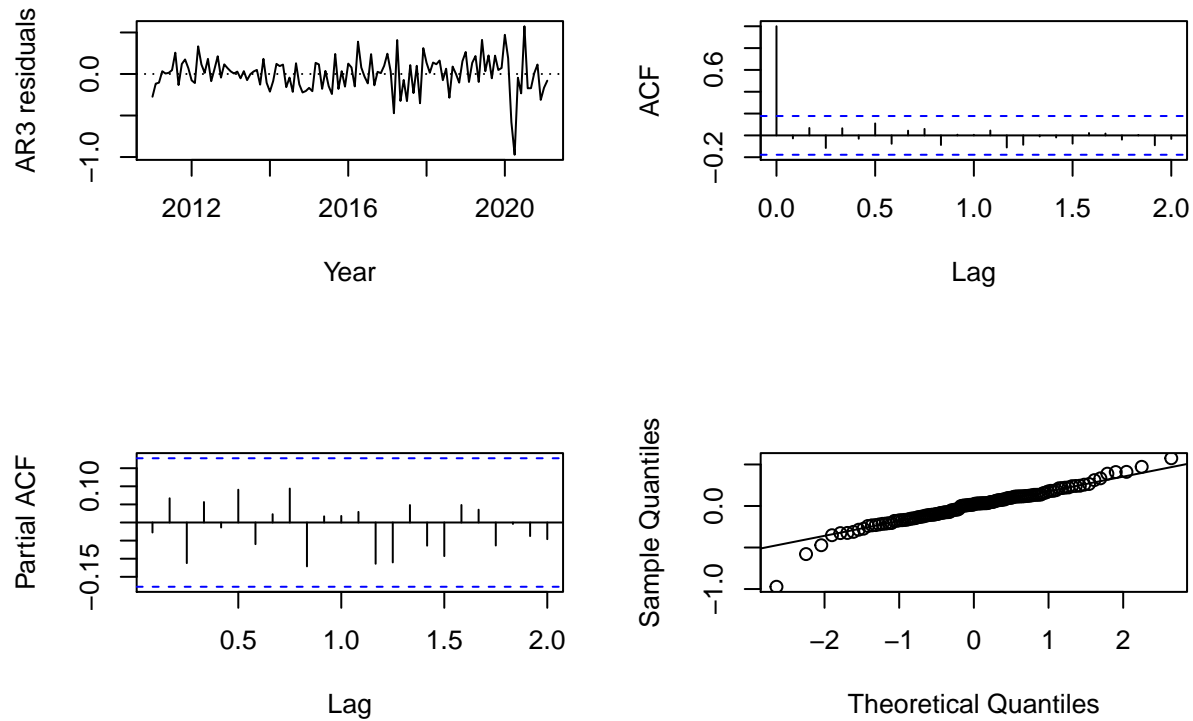

```
##
##  Box-Ljung test
##
## data:  ar3.model$residuals
## X-squared = 12.800728346589514, df = 24, p-value = 0.96929230763349972
```

For this model, the residuals do appear like they could be white noise, the acf and pacf plots do not have many lags outside of the significance threshold, the QQ-plot seems to fit well, and the p-value for the Ljung-Box test is pretty large. With this information, it is certainly possible that this series is white noise. Next I will look at the fit for the AR(4) model.

##
Box-Ljung test
##
data: ar4.model$residuals
X-squared = 11.512772707641624, df = 24, p-value = 0.98486064980363908

Again, for this model, the residuals do appear like they could be white noise, the acf and pacf plots do not have many lags outside of the significance threshold, the QQ-plot seems to fit well, and the p-value for the Ljung-Box test is pretty large. With this information, it is certainly possible that this series is white noise. Next I will compare the AIC's for the three AR models and choose which model I think is best for this data.

AR(2)   AR(3)   AR(4)
## -15.788735796512867 -28.067284091350359 -27.098854708926012

Since the AR(3) model has the lowest AIC, I will use this model. My final model is a seasonal means model with a quadratic trend plus an AR(3) stationary process. Here is the estimated model:

y_t = -288.716455657048t + 0.0727863314026671t^2 + $\hat{s}_{month}$ + x_t, x_t = 1.47015299566752x_{t-1} + -0.95465883177272x_{t-2} + 0.334317345246589x_{t-3} + w_t, $w_t \sim^{iid}$N(0,0.0427570577500484), where the estimated seasonal effects ($\hat{s}_{month}$) are $\hat{s}_{jan}$ = 286474.88725638, $\hat{s}_{feb}$ = 286475.433793661, $\hat{s}_{mar}$ = 286475.754219579, $\hat{s}_{apr}$ = 286475.679755091, $\hat{s}_{may}$ = 286475.57927968, $\hat{s}_{jun}$ = 286475.476493349, $\hat{s}_{jul}$ = 286475.313896097, $\hat{s}_{aug}$ = 286475.387987923, $\hat{s}_{sep}$ = 286475.467768828, $\hat{s}_{oct}$ = 286475.584438811, $\hat{s}_{nov}$ = 286475.205197874, $\hat{s}_{dec}$ = 286474.621146015

The seasonal effects and the quadratic trend were calculated in R using the lm() function with a factor for the month.

Next, I am going to forecast the original data out several time periods, using my model, and provide prediction intervals. Here is the original data with a two year forecast and the 95% prediction interval superimposed:

Forecast
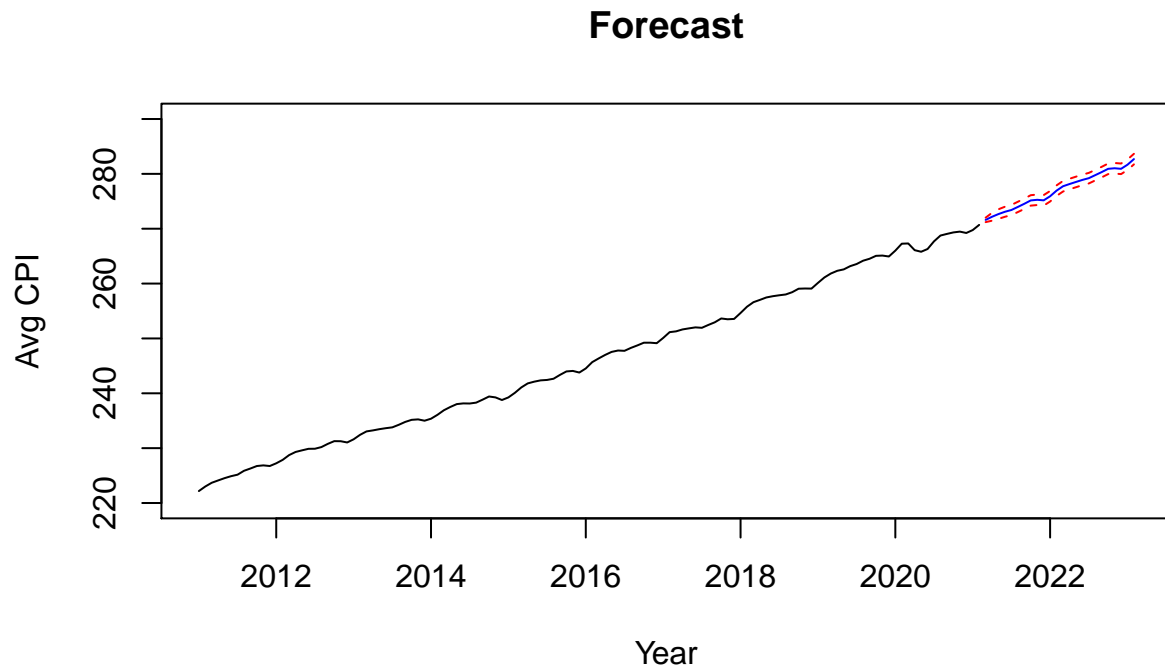
iii. SARIMA Modeling

In the next part of my analysis, I will handle the trend and seasonal components by appropriately differencing the data and fitting an appropriate SARIMA model. First I will difference the deseasonalized data and look at the resulting plots so I can fit appropriate SARIMA models.
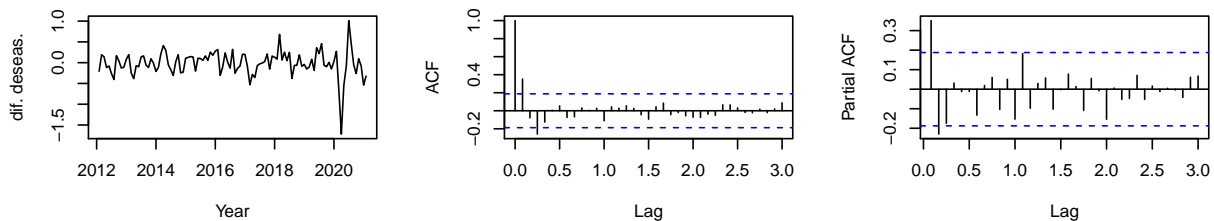

The trend appears to be removed, so I will continue with finding SARIMA models to fit. Within seasons, the acf plot seems to decay somewhat geometrically and the pacf plot appears to cut off after the first few lags. Between seasons the acf plot looks like it cuts off while the pacf plot seems like it could be geometrically declining. With this information I will try to fit ARIMA(2,1,0)x$(0,1,1)_{12}$, ARIMA(3,1,0)x$(0,1,1)_{12}$, and ARIMA(4,1,0)x$(0,1,1)_{12}$ models to the data. First I will look at the fit for the ARIMA(2,1,0)x$(0,1,1)_{12}$ model.
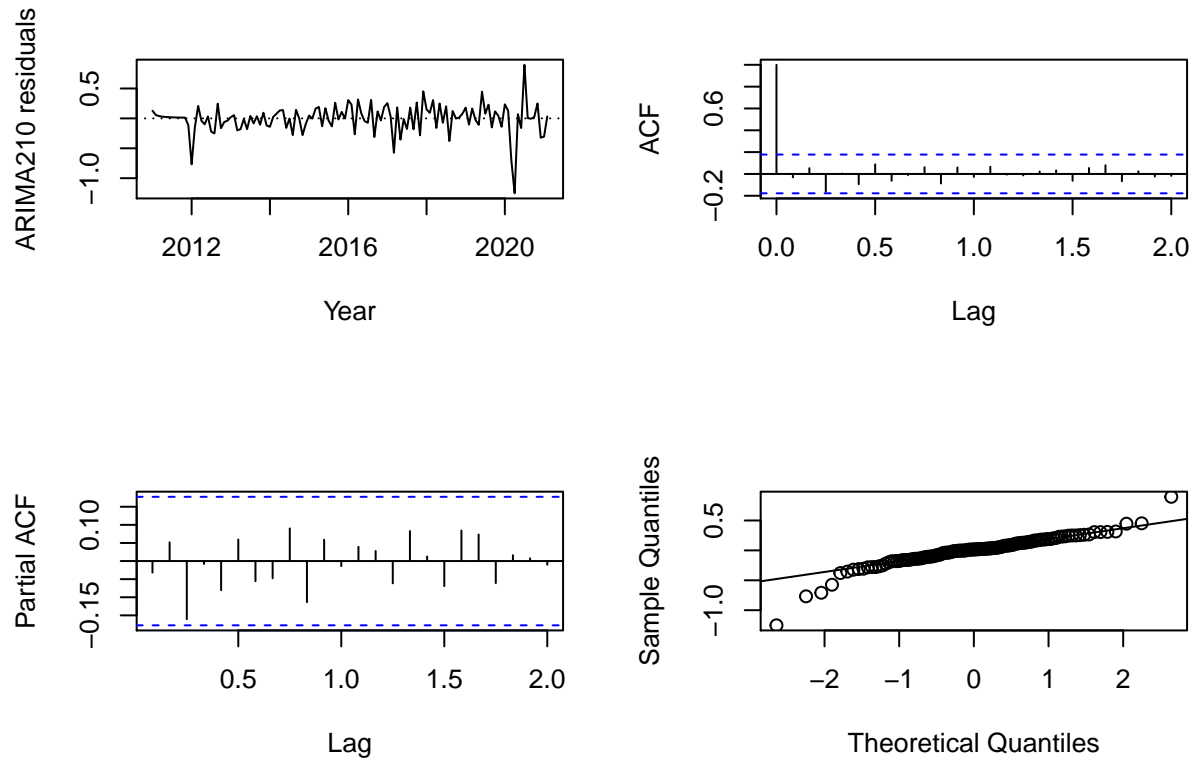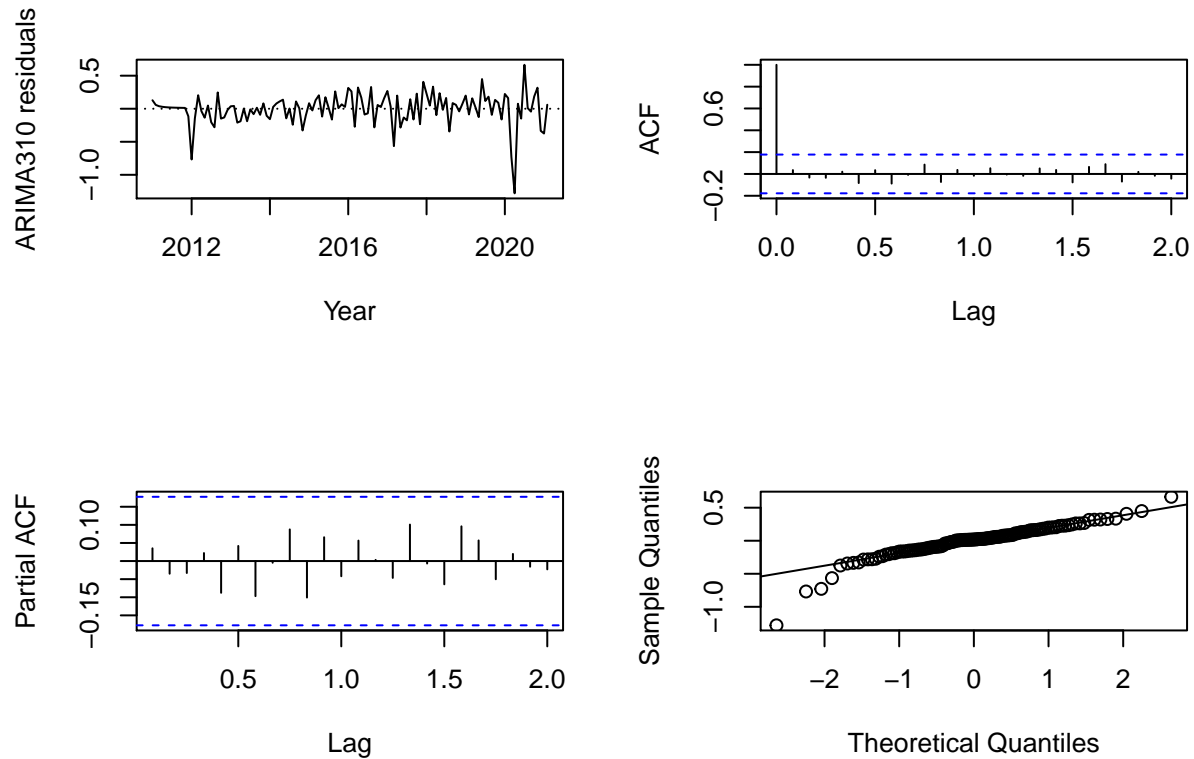
```
##
##  Box-Ljung test
##
## data:  arima210011.model$residuals
## X-squared = 12.369651080242722, df = 24, p-value = 0.97541432342642487
```

For this model, the residuals appear like they could be white noise, the acf and pacf plots do not have many lags outside of the significance threshold, the QQ-plot fits well, and the p-value for the Ljung-Box test is large. With this information, it is certainly possible that this series is white noise. Next I will look at the fit for the ARIMA$(3,1,0)x(0,1,1)_{12}$ model.
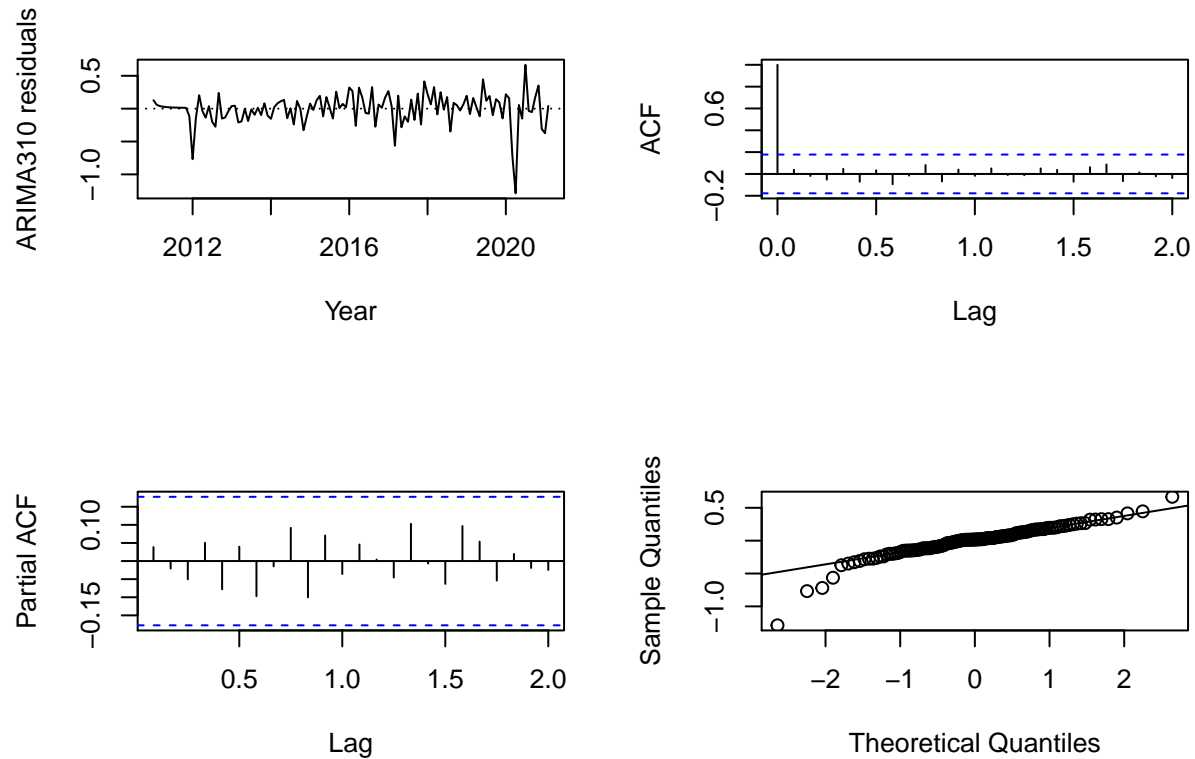
```
##
##  Box-Ljung test
##
## data:  arima310011.model$residuals
## X-squared = 9.0113935659164426, df = 24, p-value = 0.99757132297718065
```

Again, the residuals appear like they could be white noise, the acf and pacf plots do not have many lags outside of the significance threshold, the QQ-plot fits well, and the p-value for the Ljung-Box test is large. With this information, it is certainly possible that this series is white noise. Next I will look at the fit for the ARIMA$(4,1,0)x(0,1,1)_{12}$ model.

```
##
##  Box-Ljung test
##
## data:  arima410011.model$residuals
## X-squared = 9.3126488298020096, df = 24, p-value = 0.99684954101067991
```

Yet again, the residuals appear like they could be white noise, the acf and pacf plots do not have many lags outside of the significance threshold, the QQ-plot fits well, and the p-value for the Ljung-Box test is large. With this information, it is certainly possible that this series is white noise. Next I will compare the AIC's for the three SARIMA models and choose which model I think is best for this data.

```
## ARIMA(2,1,0)x(0,1,1)  ARIMA(3,1,0)x(0,1,1)  ARIMA(4,1,0)x(0,1,1)
##    18.382362921341866    16.959731039794889    18.772363322174982
```

Since the ARIMA(3,1,0)x$(0, 1, 1)_{12}$ model has the lowest AIC, I will use this model as my final model. Here is the fitted model:

$(1-(0.48714701448344\,B)-(-0.299795802876717\,B^2)-(-0.17587757679283\,B^3))\nabla(\nabla_{12}x_t)=(1+-0.671793947521319\,B^{12})w_t$, $w_t \sim$wn(0,0.058130316578913)

Next, I am going to forecast the original data out several time periods, using my model, and provide prediction intervals. Here is the original data with a two year forecast and prediction interval superimposed:
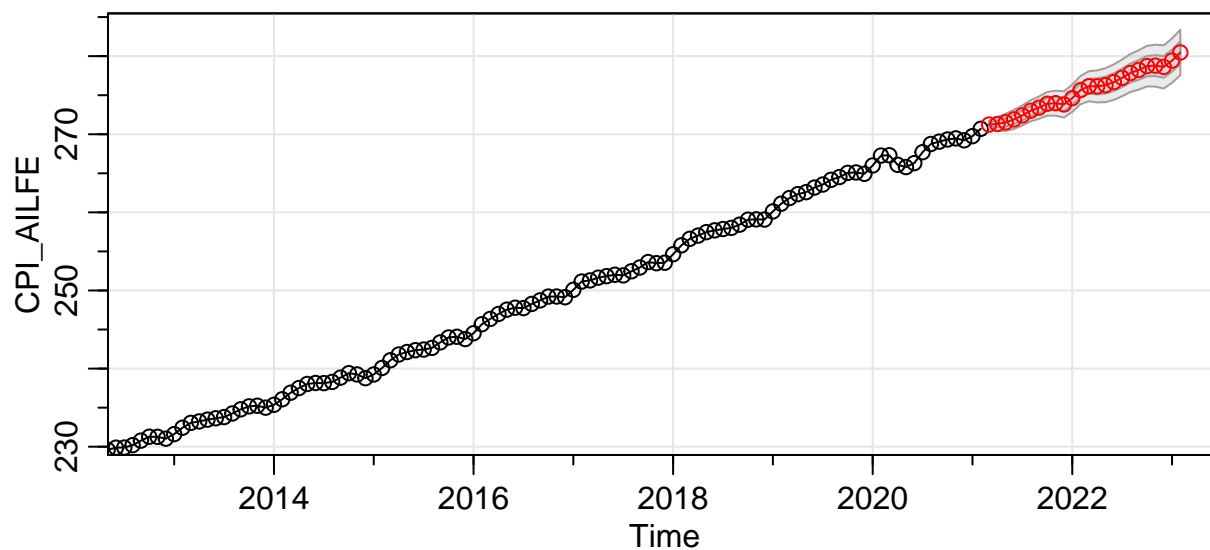
iv. Model Comparison

Now I am going to compare the models I chose in part ii nd part iii. Here are the models superimposed on the original data and their corresponding residual plots:

Seasonal means model

SARIMA model

Seasonal means model

SARIMA model
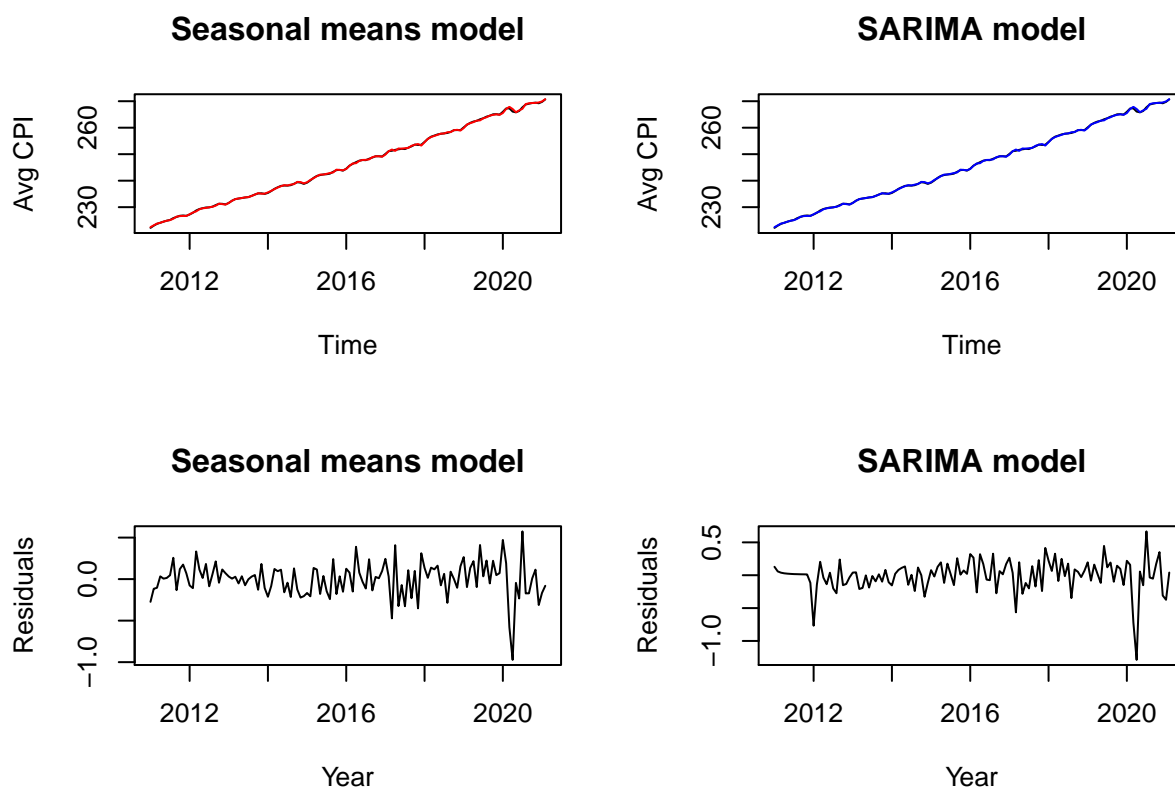
When looking at the models, we can see that the plots of the models are hard to tell apart because they fit the

data in a similar fashion. The residual plots also look similar (although not as similar as the plots of the models), but when looking closely it seems as though the residuals for the seasonal means model may be slightly smaller. Also, the AR(3) model for the detrended data has a smaller AIC than the $ARIMA(3,1,0)x(0,1,1)_1 2$ model.

v. Conclusions

Since I believe the residuals for the seasonal means model may be slightly smaller, since the AIC was smaller, and because the data seems to be dominated by a trend that I easily modeled with this method, I believe that the seasonal means model with a quadratic trend that I fit in part ii is the best model. One last thing I would like to note about this analysis, is that there seems to be a dip and spike for some of my plots during 2020. I believe this is due to the COVID-19 pandemic, which was why I was not too worried about this while I was modeling the data.