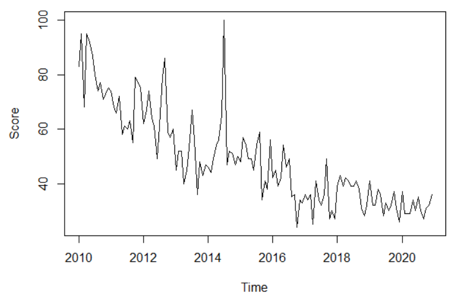
J-pop Report

Like the report on the K-pop analysis, we must first see if the data is stationary. Meaning we have to look at a plot of the data and see if there are any obvious signs of trends and/or seasonality.



Since there is a decreasing trend, this is a non stationary data. Also, there seems to be no seasonality signs present in the plot.

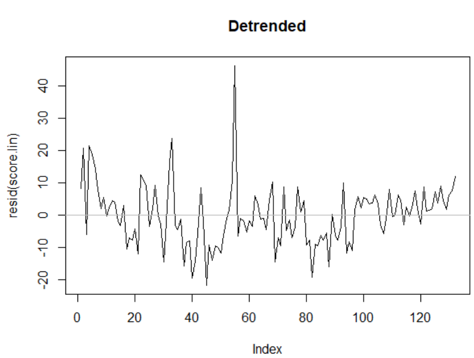
As it is not stationary, we must first make it into one. We can do many ways. However, for this, I choose to do it in two separate ways: fitting a linear model and taking the first difference. Before fitting a linear model, I performed the ADF test to see if removing the linear trend and the intercept would result in a stationary data.

```
	Augmented Dickey-Fuller Test

data:  J_pop
Dickey-Fuller = -3.9721, Lag order = 5, p-value = 0.01281
alternative hypothesis: stationary
```

Since the p-value is 0.01281, which is small, we could reject the null hypothesis that the data is not stationary after the removal. Hence, we could go through fitting a linear model.

After fitting a linear model to our data, we can see from the image below that the residual of the model shows a better stationarity than the initial data did.

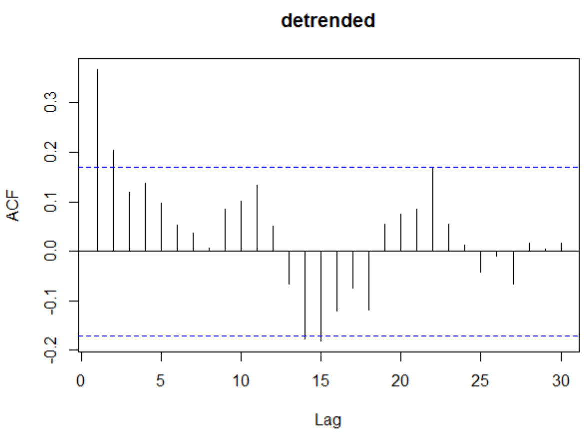


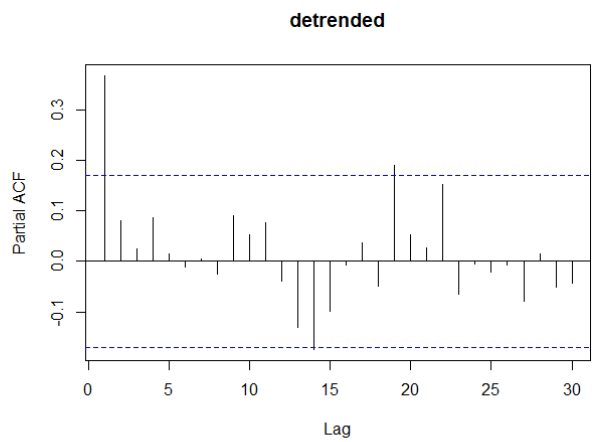
To ensure that the residuals of the linear model is stationary, we performed the ADF test again:

```
	Augmented Dickey-Fuller Test

data:  resid(score.lin)
Dickey-Fuller = -3.9721, Lag order = 5, p-value = 0.01281
alternative hypothesis: stationary
```

With a p-value of 0.01281, it is!

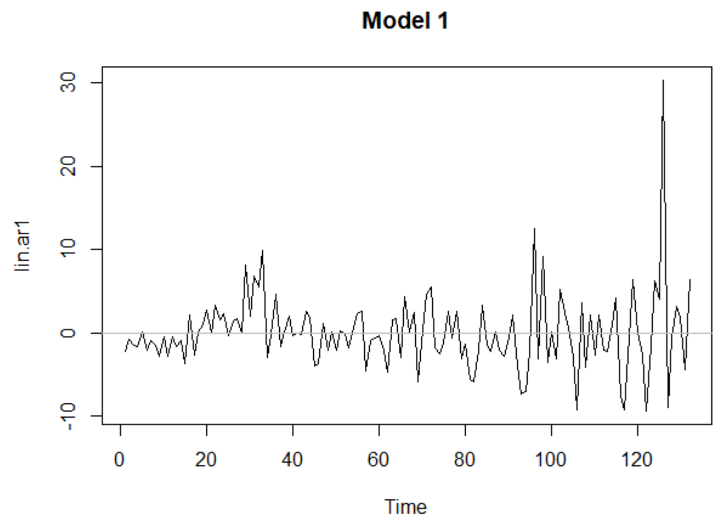


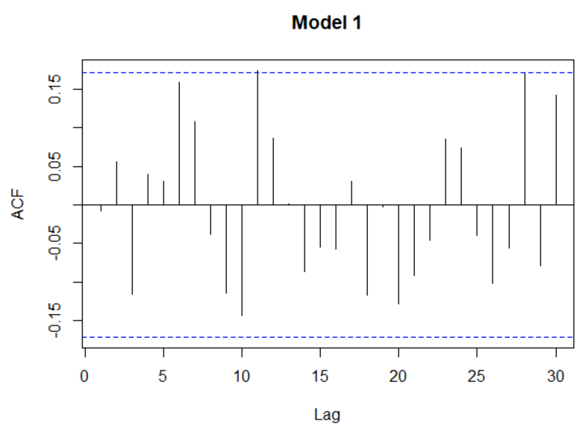


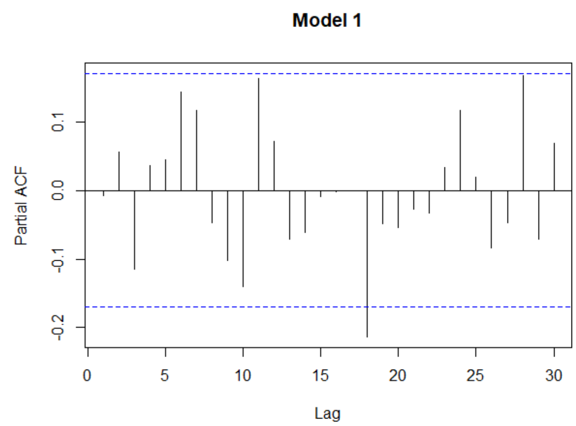
Taking the ACF and PACF plots of the residual of the linear model, we can see from the ACF plot that it gradually decreases after the first lag. The PACF plot also drops down after the first lag. Hence, this is an AR process of lag 1. To check this, I used some ARIMA models where I gave p and q a value of 1. Here are the results:

Model	AIC	BIC	MAE	RMSE	MAPE	MASE
Arima(1,0,0)	781.73	790.38	3.014773	4.564776	483.1207	0.857891
Arima(0,0,1)	791.44	800.09	3.185456	4.737945	382.2522	0.906461
Arima(1,0,1)	783.68	795.22	3.015311	4.5639	495.3109	0.858044

From the table, we can see that the AR model with lag value of 1 has the minimum AIC, BIC, MAE, MAPE, and MASE values. With the exception of RMSE value, an ARIMA(1,0,0) model is the best fit.





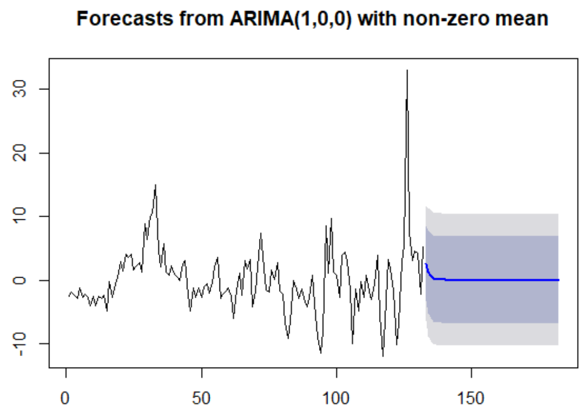


Taking the residual of our model, we can see from the lone plot above that the mean and variance seems to be constant. However, looking at the ACF and PACF plots, there seems to be significant lags. In order to ensure that the residuals are independent by nature, I performed the Box-Pierce test.

```
        Box-Ljung test

data:  lin.ar1
X-squared = 2.2637, df = 1, p-value = 0.1324
```

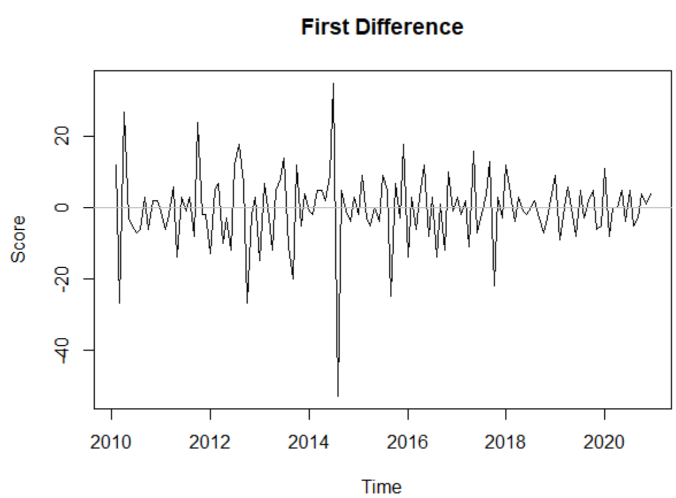
Since the p-value of the test is large, we could not reject the null hypothesis that the residuals of our model are independent. Plotting the forecast 50 months into the future, the image is below:

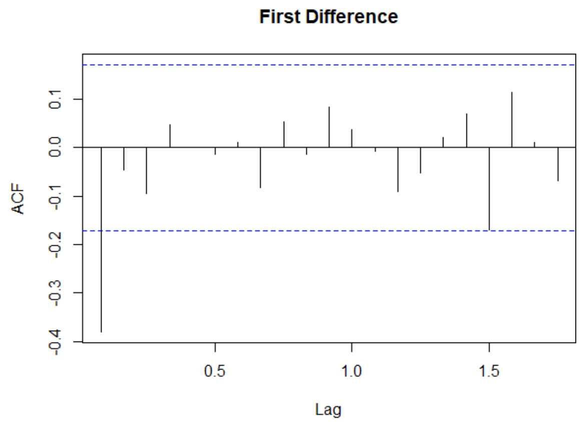


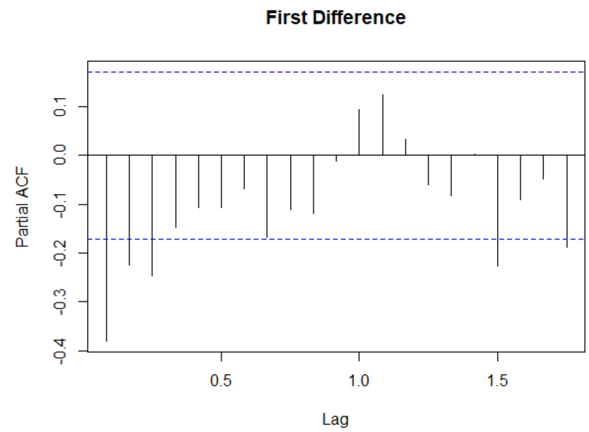
Looking at the plot of the forecast, we can only see that the scores of the data went down and stayed constant. This does not tell a lot about the possible scores of the data. Hence, we could further try more models.

Before I go do more models, I thought of analyzing the data when de-trended using first difference.

Taking the plot of differenced data, we can see that the scores have a more constant values than when fitting a linear model.



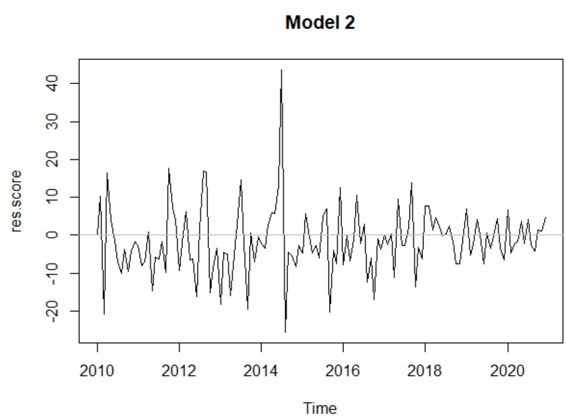


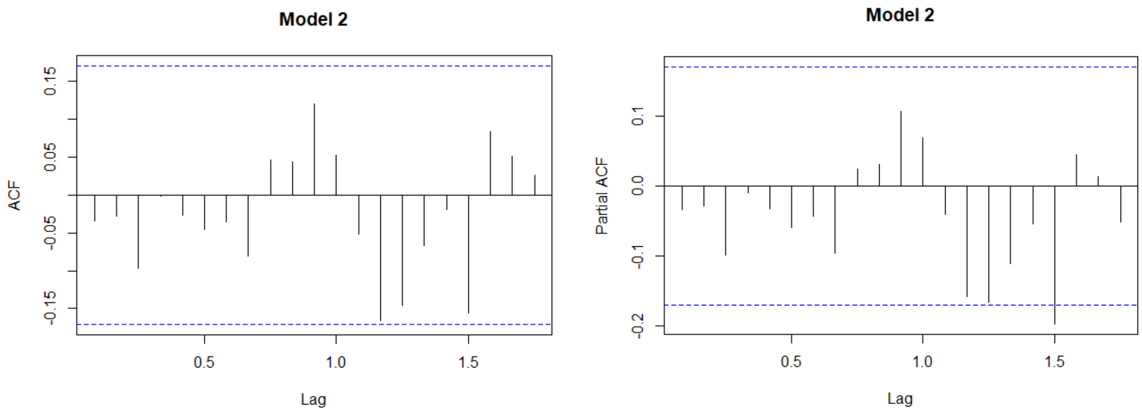


From the ACF and PACF plots, there is a sharp drop after the first lag in the ACF plot and a gradual decrease in the PACF plot. This shows that there might be a moving average process in the differenced data. To check this, I tried several models:

Model	AIC	BIC	MAE	RMSE	MAPE	MASE
Arima(1,0,0)	989.24	992.12	7.190023	10.43704	15.50425	0.695248
Arima(0,0,1)	970.95	976.70	6.710686	9.653433	14.22153	0.6488979
Arima(1,0,1)	953.86	959.61	6.570049	9.026869	13.87013	0.6352988
Arima(1,1,1)	952.94	961.56	6.451227	8.925526	13.75307	0.6238092

We can see from the table above that the best model in this case is ARIMA(1,1,1) as it has the lowest values for AIC, MAE, RMSE, MAPE, and MASE. Though is has a slightly bigger BIC value than another model, it fulfills most criteria, hence, we will go with this.



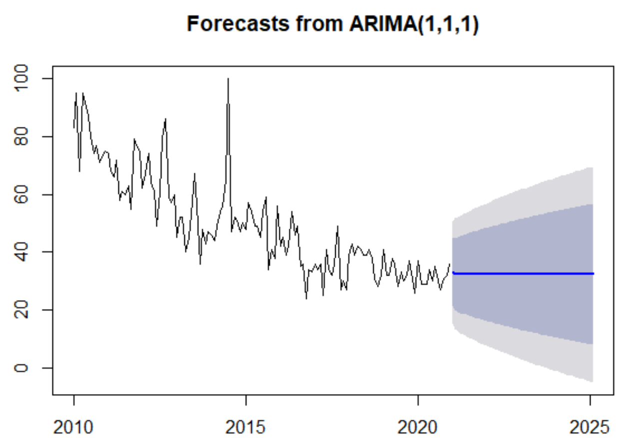


Looking at the plots above, there seems to be no significant lags with the exception at PACF plot. Nonetheless, I performed Box-Pierce test to make sure the residuals of this model are uncorrelated.

```
	Box-Pierce test

data:  res.score
X-squared = 1.4872, df = 1, p-value = 0.2227
```

Since the p-value of our test is large, we could not reject the null hypothesis that the residuals of our model are independent.

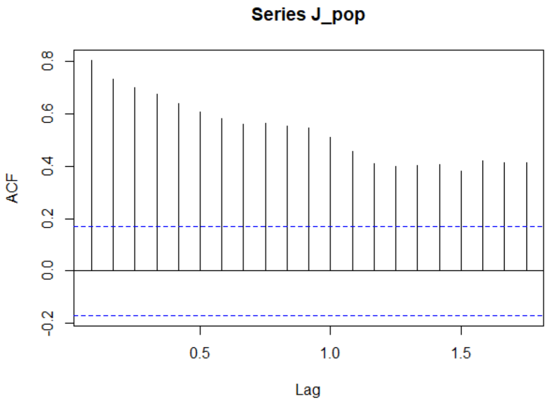


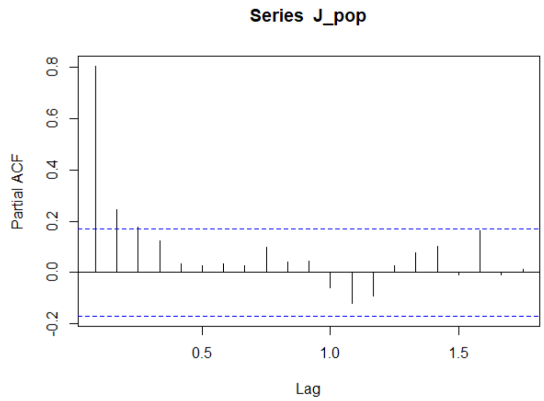
Like the forecast to the first model of this project, this forecast also does not tell a lot about possible scores in the future.

What is surprising doing both de-trending ways is that both gave two different models. Fitting a linear model, we got an autoregressive process model while taking the first difference we got a moving average model. The only thing both models have in common is that both are significant at lag 1.

Seasonal Models

I was going to try seasonal models as well since I'm already doing some models. However, I am also aware that there is no evidence that suggests seasonality exists in this data. Not only does the plot of the data shows no evidence, but also the ACF and PACF plots shows none.





The ACF shows that there are a lot of significant spikes, not just one that repeats at a certain interval. The PACF plot also shows no exponential decay at its seasonal lags. I could not reason out why would there be a seasonal component in my data, hence, I would not perform this function.

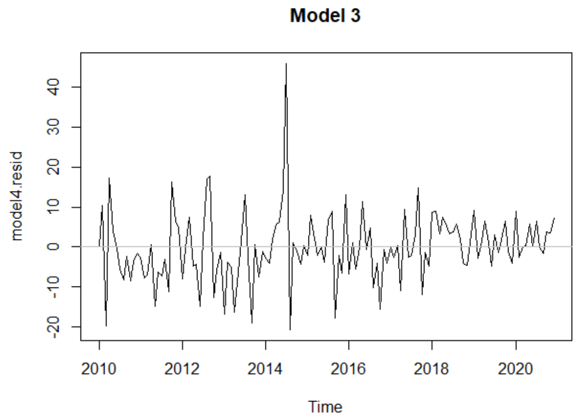
Drift

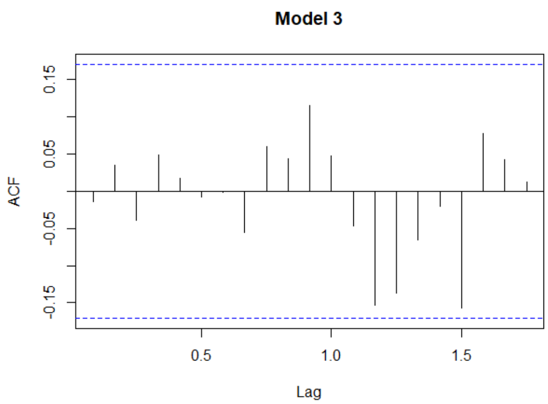
My third possible model would include a drift using the autoregressive model Arima(1,0,0). While the fourth possible model would also include a drift using the moving average model Arima(1,1,1).

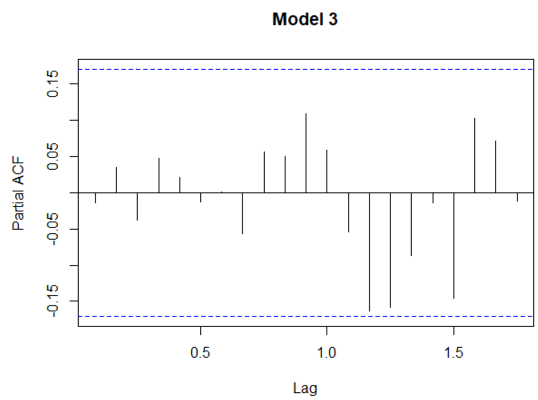
These models, gave the following summary:

	AIC	BIC	MAE	RMSE	MAPE	MASE
Model 3	953.13	964.66	6.233516	8.67576	13.3563	0.6027574
Model 4	948.82	960.32	6.174436	8.700671	12.94106	0.5970446

Looking at the results above, the model Arima(1,1,1) with drift gave the best values.



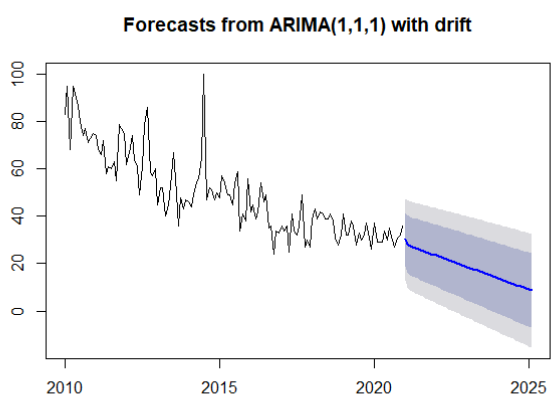




Since the ACF and PACF plots shows no significant lags, we can say that the residuals are not correlated. Even performing a Box Test function shows the same result:

```
        Box-Pierce test

data:  model4.resid
X-squared = 0.74164, df = 1, p-value = 0.3891
```



The plot of the forecast shows a better result than the two previous models. As seen the scores decreases rather than going constant.

Auto Arima

Performing the auto Arima function, I got the same result as my third possible model. That is ARIMA(1,1,1) with drift.