Synopsis

With the market being volatile and inflation being unpredictable at times, I was interested in finding what the trend of a government worker's payment would be. Talking with family and friends working in the government, they were stress-free about their income and job-stability as their pay does not get disturbed by the market and follows inflation. This was a topic of interest as the salary of a public servant would be a good benchmark to measure one's compensation increase (assuming same position/seniority).

Another reason this was a topic of interest was that one can know if their pay is following the trend of the current economy and government worker's payment would be a good to use as a comparison. Knowing the public sector salary is also important since they operate based on taxpayer money unlike corporations that are financed via private investors and revenue from capitalistic policies. Thus, it is crucial they are open and more accountable from taxpayers (Canadian Citizens) so that the general public knows that their tax money isn't being mishandled and used in corrupt activities. Furthermore, it allows the comparing the performance of the federal government organizations with the compensation of its employees and using it to evaluate the efficiency of the government organization to other government

Background

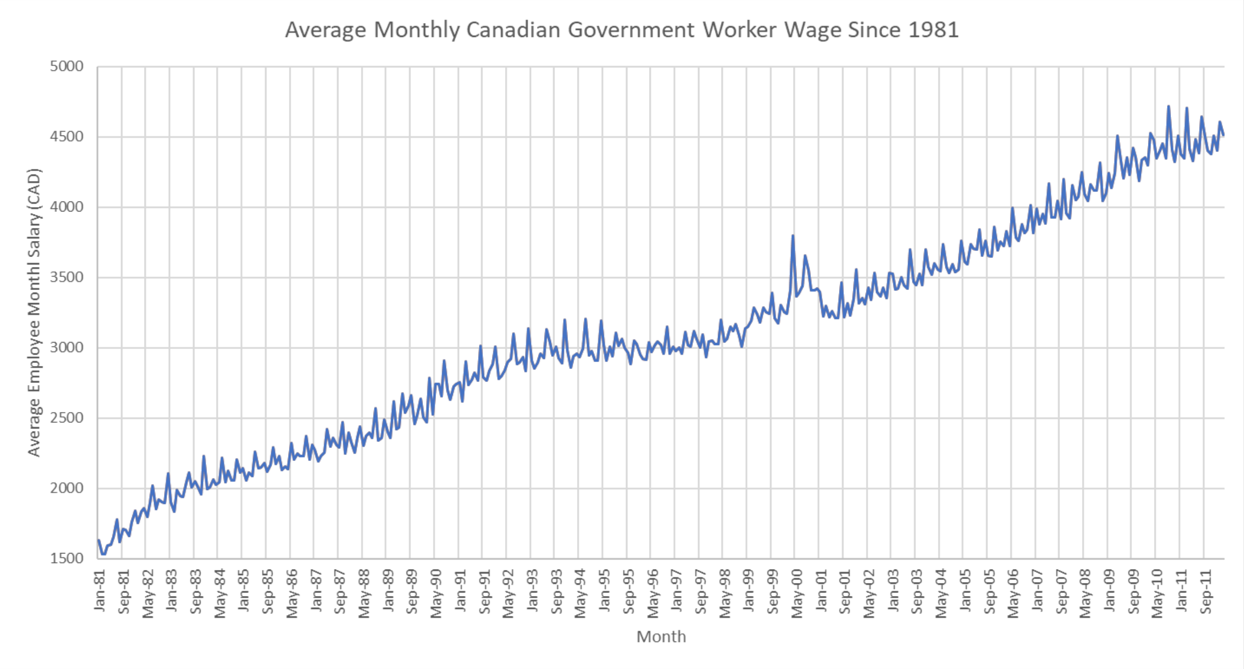
The reason behind why the Canadian Government started the project was to obtain data in employment. The data was collected by the Canadian Government via a survey and data collection from central payment system. The survey is a census with a cross-sectional design and data itself was collected for all the target population, thus meaning that no sampling was done. The data population is drawn from all institutes controlled and financed by the Canadian Government of all levels (federal, provincial, territorial, municipal). This includes each individual government departments, ministries, agencies, funds, crown corporations, public health institutions, and educational institutions.

For the institutions that use a central pay system such as Phoenix Pay System, the data were obtained by referencing the system. For the entities not using such system, a questionnaire was developed as used for search purposes. One questionnaire was used for each of the federal institutional entities and provincial/territorial entities. The survey was also sent to some government business enterprises with no database. These questionnaires were designed to gather as much information on employment and salary.

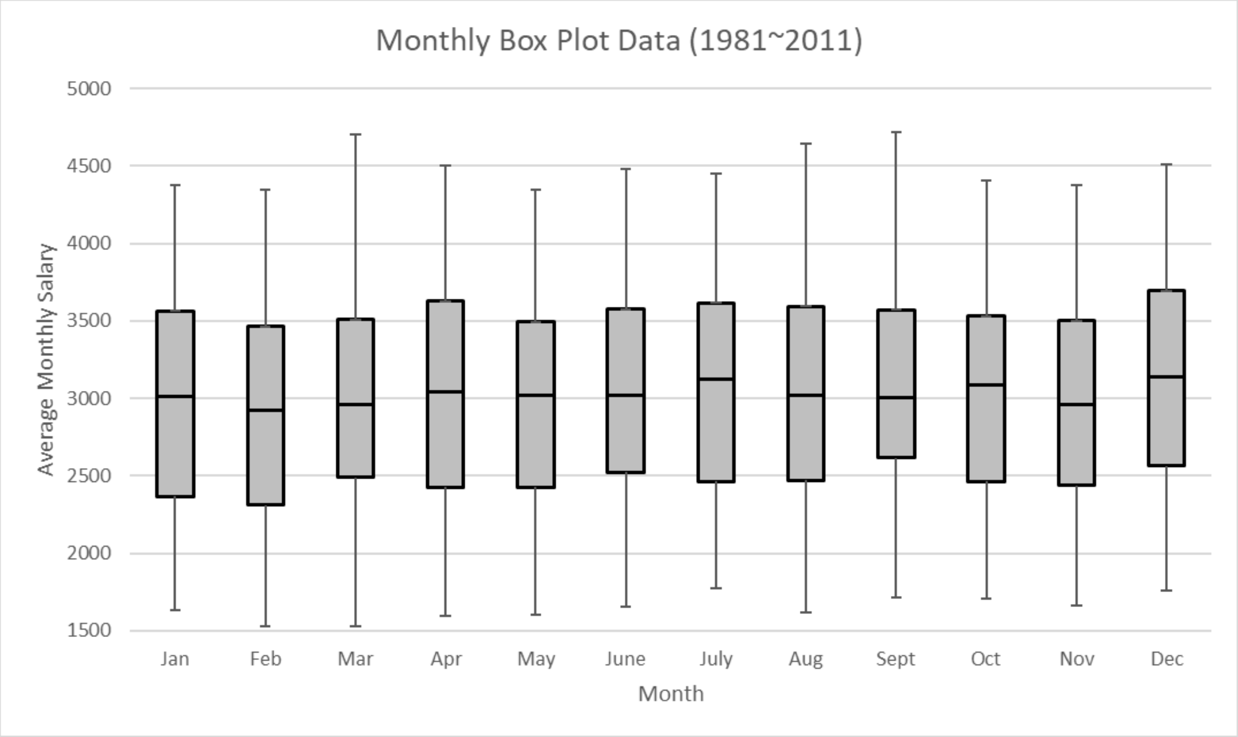
According to the statement released along with the data, the fluctuation in the time series data is due to seasonal, cyclic, and irregular fluctuations. The seasonal variation can be explained by common holiday and vacation period for workers, climate/weather affect crop yields, production and retail affected due to holidays such as Christmas. Thus, this would need to be accounted for in the data analysis.

Dataset

The times series plot of the average monthly Canadian Government Worker Wage can be seen below. The figure shows the plot for thirty years' worth of data. As there is a clear upwards trend, it can be said that the statistical property of the time series (mean, variance, etc.) are not constant over time. Therefore, there is a need to transform this data to obtain stationarity.



It can be seen clearly that there is seasonality involved with the data as the yearly highs generally happen around winter and summertime. There are two peaks in a year with one happening in the summer and one in the winter. Although, there are some discrepancies as some years the peak happen in fall and spring, these only happen in the minority of the occurrences while in majority of the years, the peaks are predictable. Below shows the Box-and-Whisker graph that compares each month that can be used to further study each month.



For each month, the whiskers are the same length with the median splitting the box approximately in half. This implies the data in homoscedastic and normally distributed. Thus, box-cox transformation in not required to transform the data. As there is no other series or inputs for the data, the cross-correlation function is not required.

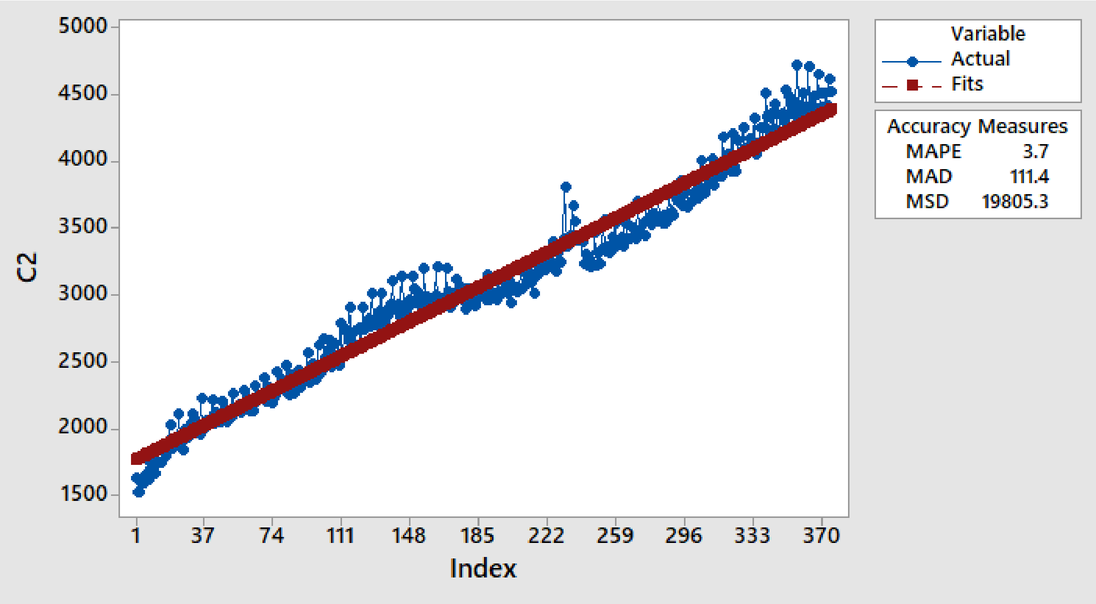
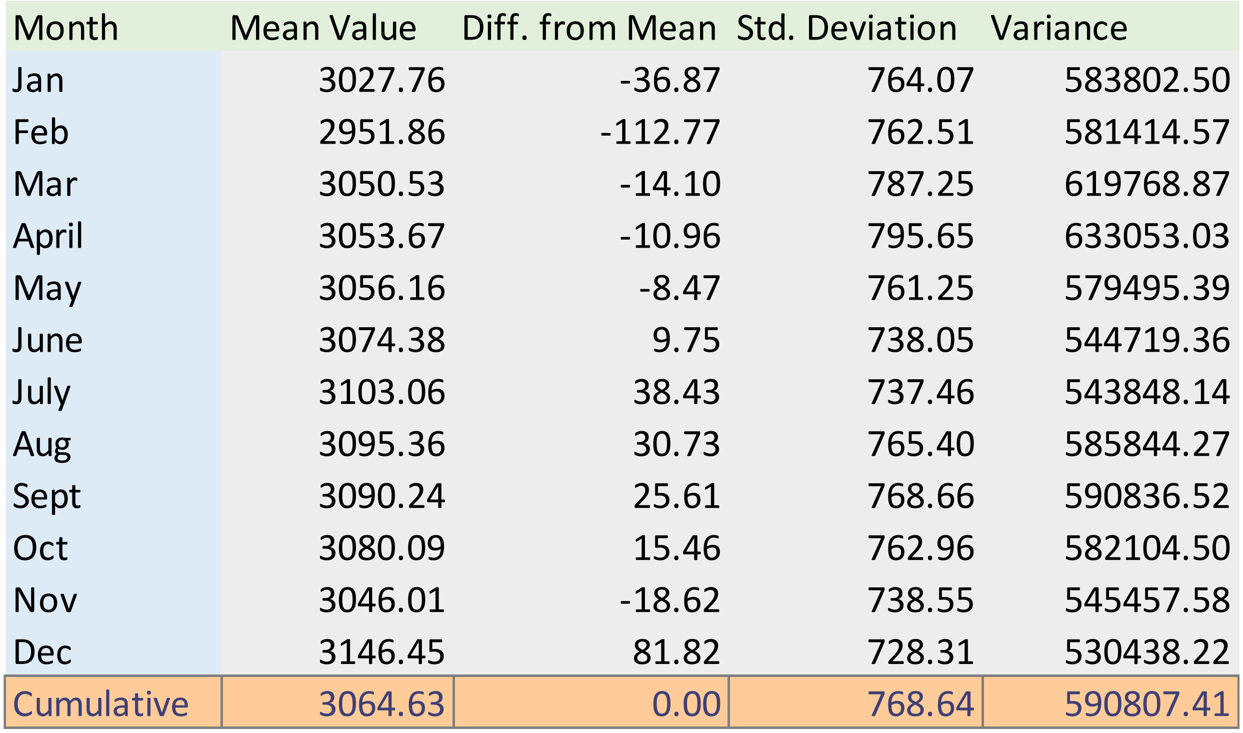
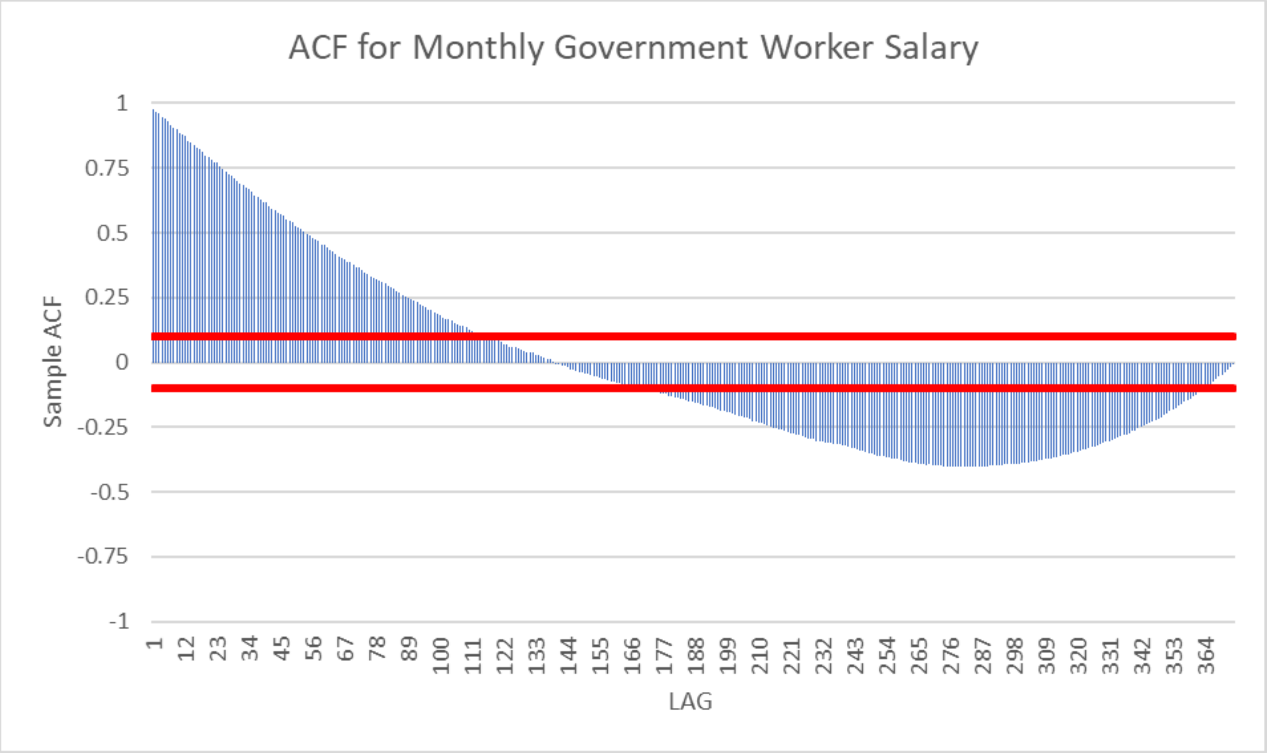


Figure above shows the overall trendline of the data. As the graph of the given time series does not seem to blur any statistical information to be shown and since seasonality and non-stationary nature of the graph can already be seen, Tukey Smoothing will not be needed for our analysis. Each season's mean, std. dev, and variance was obtained.

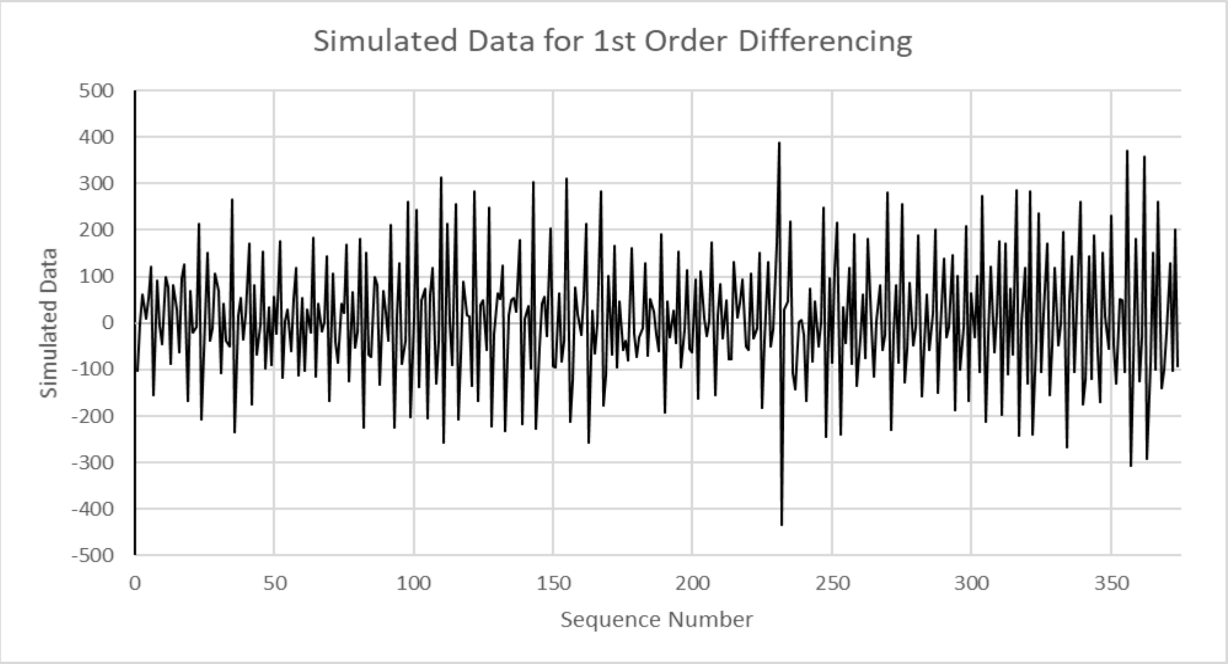


Month	Mean Value	Diff. from Mean	Std. Deviation	Variance
Jan	3027.76	-36.87	764.07	583802.50
Feb	2951.86	-112.77	762.51	581414.57
Mar	3050.53	-14.10	787.25	619768.87
April	3053.67	-10.96	795.65	633053.03
May	3056.16	-8.47	761.25	579495.39
June	3074.38	9.75	738.05	544719.36
July	3103.06	38.43	737.46	543848.14
Aug	3095.36	30.73	765.40	585844.27
Sept	3090.24	25.61	768.66	590836.52
Oct	3080.09	15.46	762.96	582104.50
Nov	3046.01	-18.62	738.55	545457.58
Dec	3146.45	81.82	728.31	530438.22
Cumulative	3064.63	0.00	768.64	590807.41

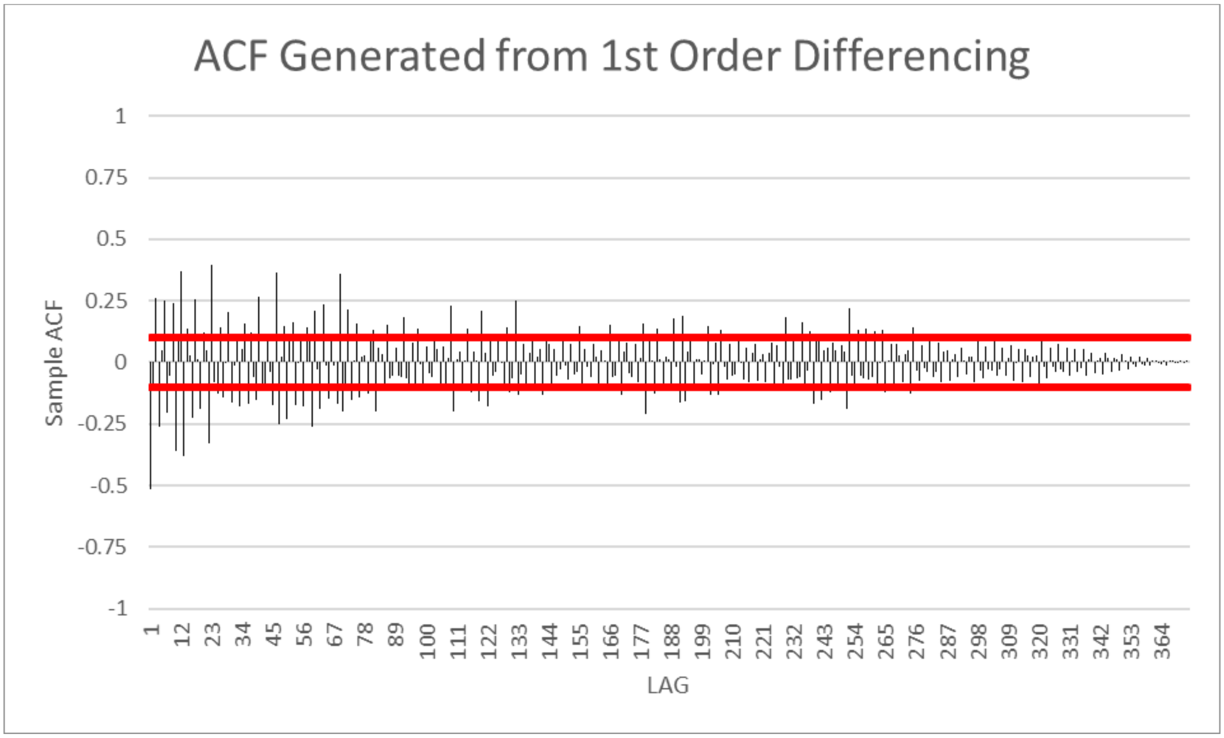
Looking at this data, the seasonality aspects are further noted. Summer months have a higher than average values along with December. The standard deviation value is used to obtain the ACF and generate the error bar for 95% confidence interval.



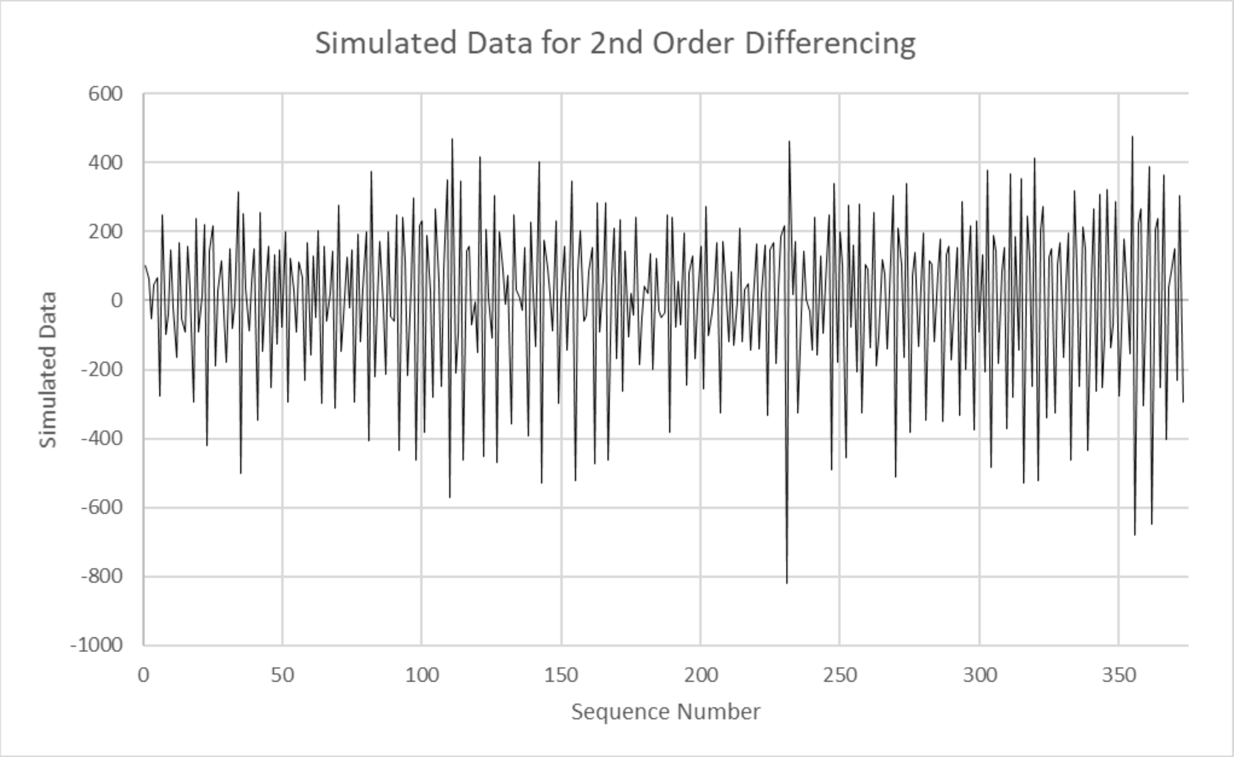
From the figure, the ACF seems to die off even though it is quite slow in dying off. This indicates that differencing is strongly suggested. The prior graphs strongly indicated that the data is non-stationary just from a visual observation. The ACF further supports the notion that differencing is required. 1st order differencing and 2nd order differencing was conducted and their respective ACFs were generated.

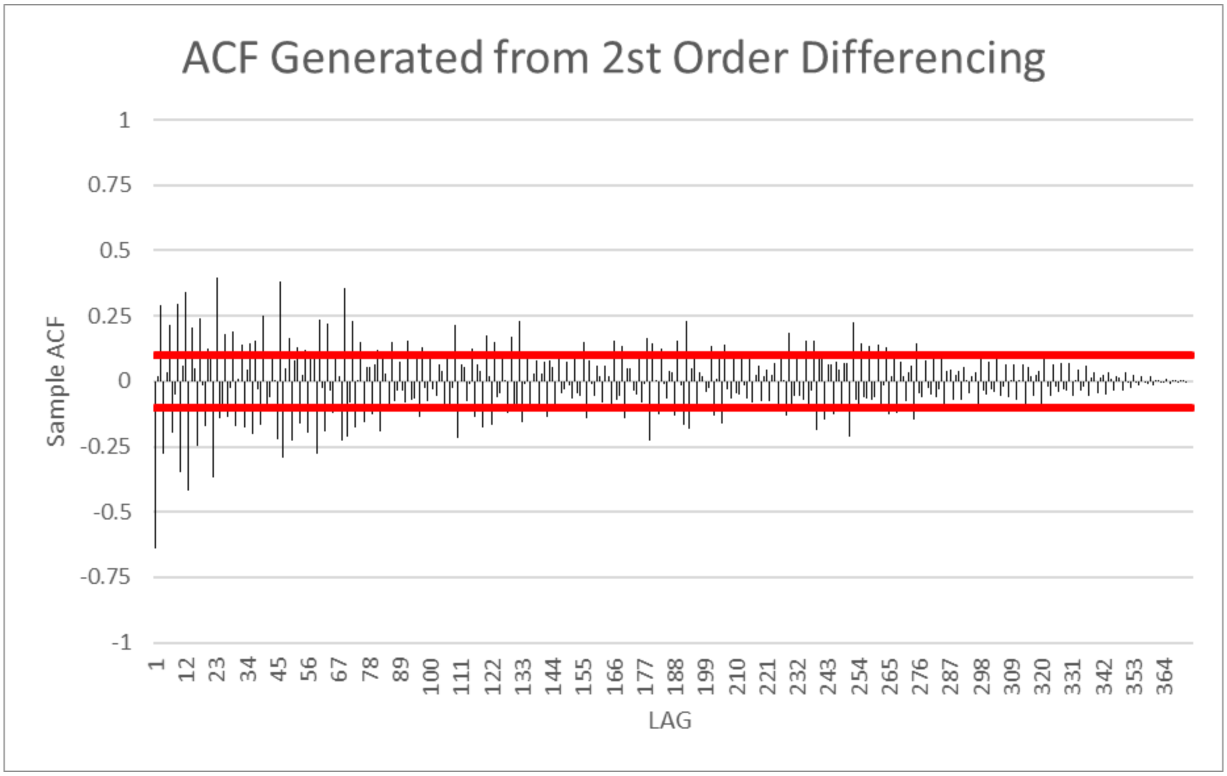


Simulated Data for 1st Order Differencing



ACF Generated from 1st Order Differencing





The figures show us that there isn't much difference whether the data was differenced once or twice. The amount of lag that falls outside of the error bar does not decrease with the additional differencing. Therefore, the differencing will be done once in the base ARIMA model as ARIMA (P, 1, Q). The differencing also has made the data stationary compared to the original graph. Even though with the non-seasonal differencing, a lot of the values are above the significant levels. Thus, normal ARIMA model will not be enough by itself, and seasonality will have to be accounted for. Earlier, it was determined that the Box-Cox transformation would not be necessary so the value of λ is set as 1. With this, other parameters in the model can now be determined. SARIMA model looks like following:

$$(p, d, q)\ X\ (P, D, Q)_S$$

$$(p, 1, q)\ X\ (P, 1, Q)_{12}$$

Next, the PACF, IACF, IPACF of the differenced data was obtained and the following conclusions were made:

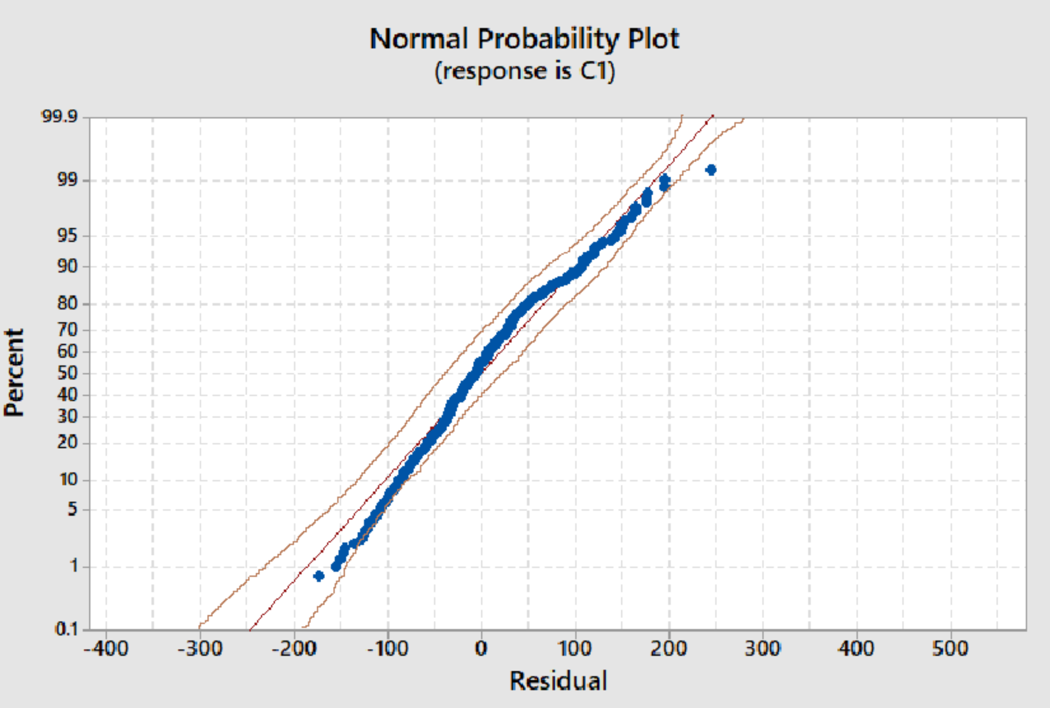
- There are significantly large values of sample ACF and IPACF at sample at lag 12 and few other seasonal spikes, seasonal MA parameter is going to be needed.
- Sample IACF attenuates after first few lags, non-seasonal MA is needed.
- ACF seems to get cut off, non-seasonal AR parameter may not be needed. However, the IPACF has a large spike at lag 1 followed by a decreasing wave that fluctuates between the positive and negative values meaning that non-seasonal AR parameters may be required.
- Seasonal spikes in IPACF attenuates, seasonal AR is required.

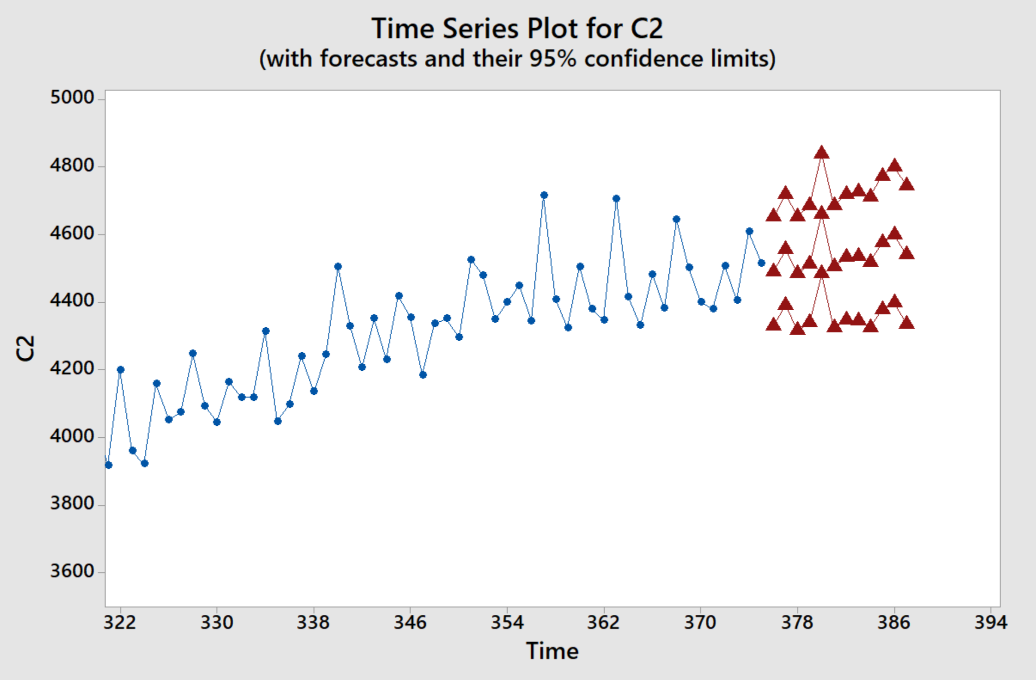
The maximum likelihood estimations were obtained via Minitab software and was generated for each set of parameters. K value was generated for each combination by adding up the number of parameters.

SARIMA (p, 1, q) x (P, 1, Q)12				AIC
p	q	p	q	
0	3	2	3	95.384
0	4	2	2	81.690
0	4	1	1	76.567
0	5	0	3	81.690
1	2	1	2	90.500
2	2	1	2	83.880
2	2	0	2	83.626
2	4	1	1	87.380
4	2	1	1	85.377
2	4	2	2	91.920
4	4	2	2	80.445
4	4	1	1	80.567
4	4	1	2	82.760

Summary of Dataset

From looking at the AIC, $(4, 1, 4)\ X\ (2, 1, 2)_{12}$ model was chosen as the most optimal to use as it had the lowest AIC value. The parameters obtained using maximum likelihood estimations as well. RACF of this model was graphed below. Along with the forecast of the model with their 95% confidence intervals.





From the confirmatory data analysis, various parameters of the models were checked with the AIC criterion and it was found that the ideal SARIMA model for this data set would be a SARIMA $(4, 1, 4)\ X\ (2, 1, 2)_{12}$. The model's residual RACF was then graphed, and it was found that there was no significant correlation detected. Meaning the residuals are independent and white. The residuals also passed the normality testing by using the normal probability plot of the residuals. The model was also assumed to have passed the constant variance test and was ready for the forecasting and simulations phase.

The trend forecast can now be used to compare one’s own salary projection to see how they match up to a government worker’s projection, who in theory, should not be getting salary bumps other than reasons of inflation and rise in cost of living. Thus, if one’s salary projection is less than that of a government worker’s, it might indicate that they are not getting the compensation they deserve.