FAITH HA &
CHAU LE

2024

TIME SERIES
EMPLOYMENT

Prediction

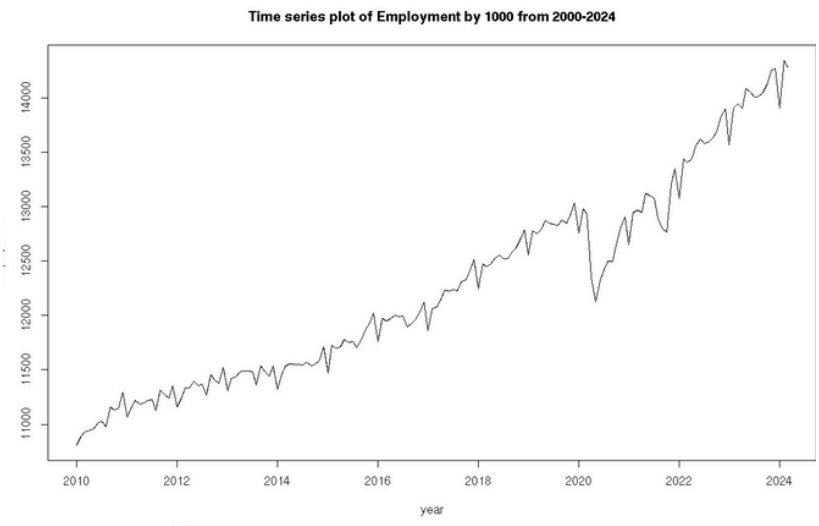

Figure 1: Graph of Employment Dataset

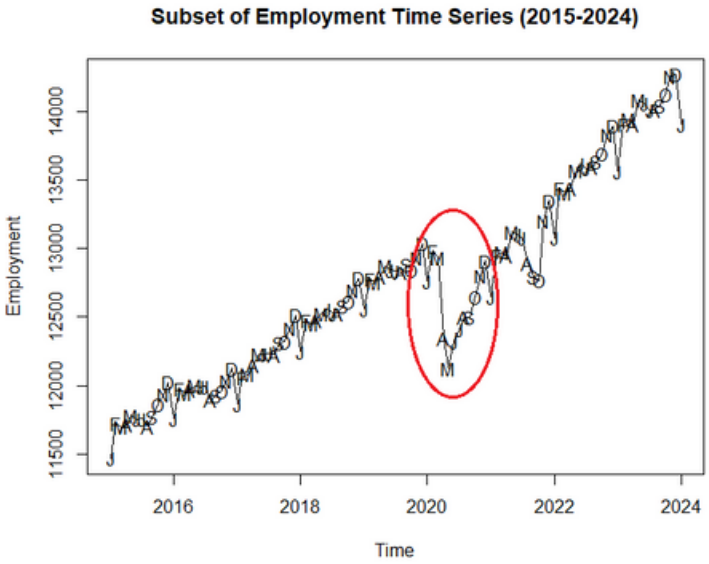

Figure 2: Subset of Employment Dataset

DESCRIPTIVE ANALYSIS

Employment is a crucial aspect of Australia's economy. To accurately forecast employment numbers in Australia, we will use time series forecasting techniques. These methods allow us to analyze historical employment data, identify patterns and trends, and make informed predictions about future employment levels. By employing sophisticated statistical models, we aim to provide reliable forecasts that can aid policymakers, businesses, and other stakeholders in making data-driven decisions to support economic growth and stability.

Our research question is "What are the next 10 month employment numbers for Australia?". Figure 1 illustrates the employment numbers per 1000 from January 2010 to March 2024. This specific dataset was sourced from the Australian Bureau of Statistics and was extracted from a larger dataset spanning from February 1978 to March 2024. The decision to extract a subset of data was made due to extended model run times. The graph displays an upward trend in the dataset and indicates the presence of seasonality. To further investigate, the dataset was subsetted, revealing ongoing seasonality, as depicted in Figure 2. The time series also demonstrates an autoregressive structure attributed to the seasonality. Notably, a significant change point occurred during the onset of COVID-19, highlighted in Figure 2.

SUMMARY STATISTICS

Min	1st Qu.	Median	Mean	3rd Qu.	Max
10810	11474	12080	12254	12859	14345

Table 1: Summary Statistics

Table 1 shows the summary statistics for the time series, the minimum employment number is 10810, the employment mean is 12254, and the maximum employment number is 14345.

NORMALITY

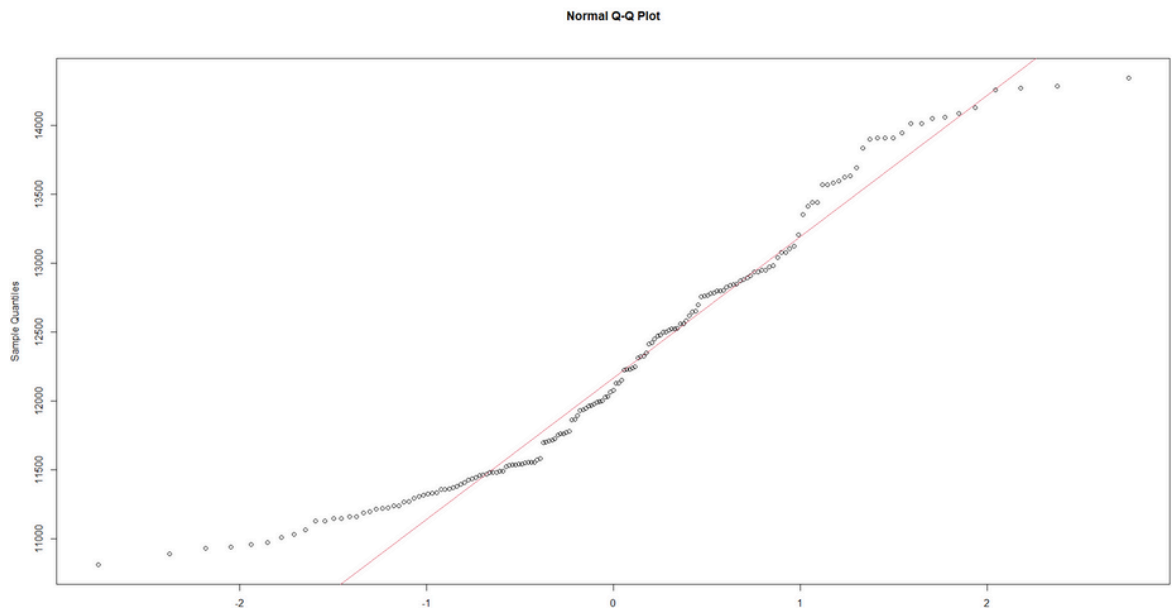

Figure 3: QQ Plot of Time Series

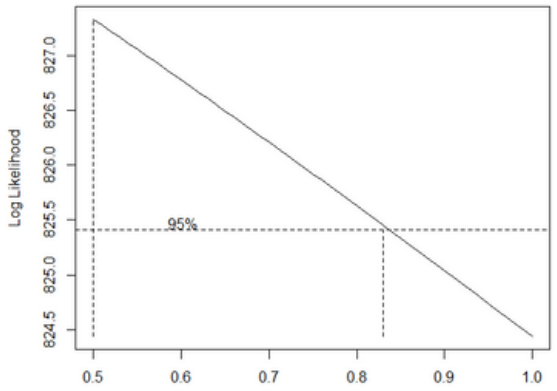

Figure 4: BoxCox Graph

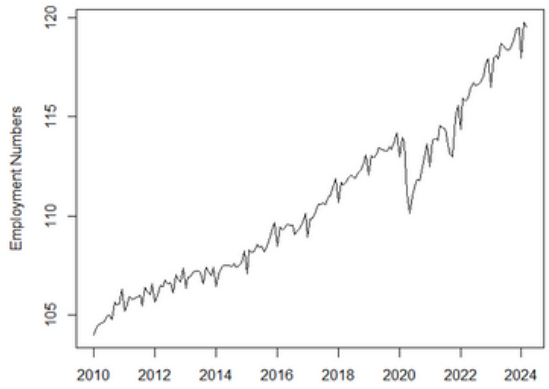

Figure 5: Square Root Transformed Data

Box Cox Transformation

As depicted in Figures 3 & 6, the series shows deviations from normalization, with the ends differing notably from the red line. To address this, the Box Cox transformation was applied to normalize the time series, as shown in Figure 4. The transformation resulted in a lambda value of 0.5, indicating that the series should be squared. Despite this adjustment, as shown in Figure 5 the time series did not exhibit substantial changes, and further ADF, PP, and KPSS tests confirmed that the data remained non-stationary.

Normalization Techniques

Various normalization techniques, such as log transformation and min-max standard scaling, were applied in an attempt to improve the data's statistical properties. However, these methods did not result in substantial improvements (Figures 7 & 8). Consequently, the decision was made to retain the original, unnormalized series to preserve its intrinsic characteristics and use SARIMA models to account for seasonality and trends. This approach ensures a more accurate analysis and reliable insights into employment patterns, reflecting the true nature of the data.

```
> shapiro.test(subset)

    Shapiro-Wilk normality test

data:  subset
W = 0.94173, p-value = 1.865e-06
```

Figure 6: Shapir-Wilk Test - TS

```
> shapiro.test(log_employment)

    Shapiro-Wilk normality test

data:  log_employment
W = 0.94984, p-value = 9.189e-06
```

Figure 7: Shapir-Wilk Test - log transformed

```
> shapiro.test(normalized_ts)

    Shapiro-Wilk normality test

data:  normalized_ts
W = 0.94173, p-value = 1.865e-06
```

Figure 8: Shapir-Wilk Test - normalized

STATIONARITY

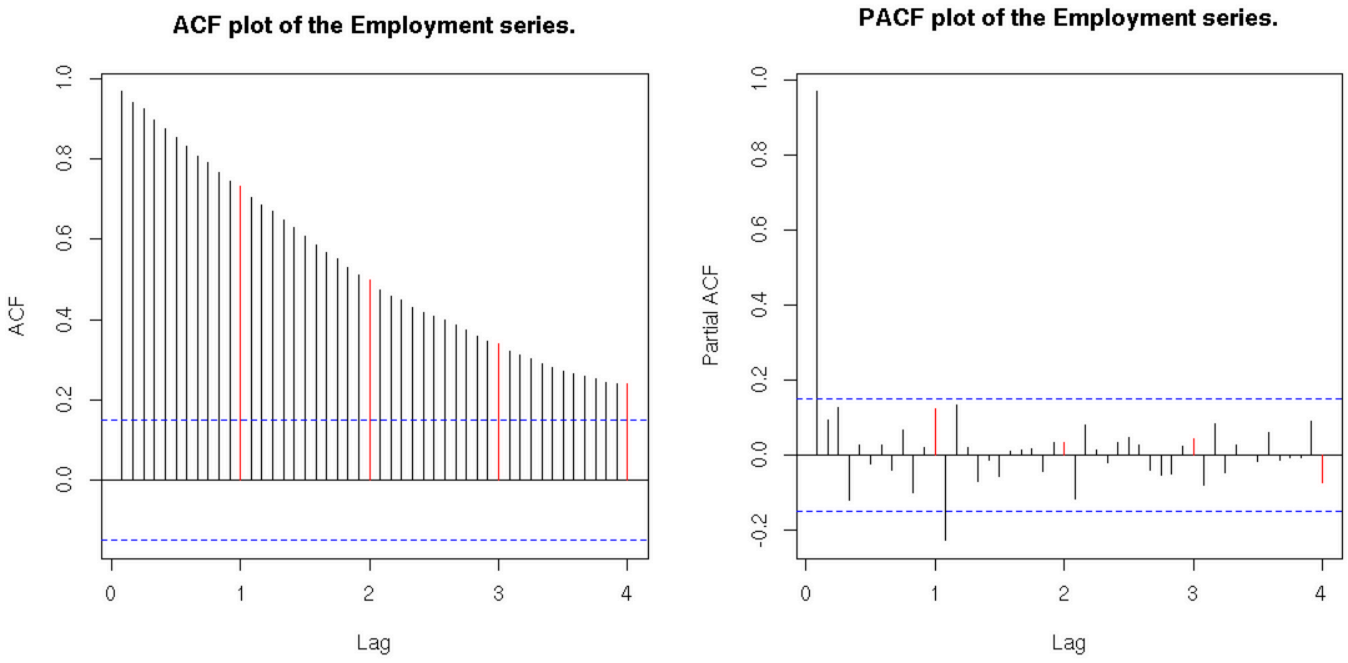

Figure 9: ACF and PACF for Employment Series

```
> adf.test(subset, alternative = c("stationary"))

    Augmented Dickey-Fuller Test

data:  subset
Dickey-Fuller = -1.4788, Lag order = 5, p-value
= 0.7939
alternative hypothesis: stationary
```

Figure 10: ADF Test for Employment Time Series

```
> pp.test(subset)

    Phillips-Perron Unit Root Test

data:  subset
Dickey-Fuller Z(alpha) = -20.242, Truncation lag
parameter = 4, p-value = 0.06136
alternative hypothesis: stationary
```

Figure 11: PP Test for Employment Time Series

```
> kpss.test(subset)

    KPSS Test for Level Stationarity

data:  subset
KPSS Level = 3.332, Truncation lag parameter =
4, p-value = 0.01
```

Figure 12: KPSS Test for Employment Time Series

Stationarity Tests

The ACF and PACF plots in Figure 9 reveal significant autocorrelation, with the ACF showing a gradual decline and the PACF displaying a notable peak. Furthermore, the ADF test indicates that the series is nonstationary, as the p-value significantly exceeds the null hypothesis (Figure 10). Similarly, the PP test supports nonstationarity, as its p-value surpasses the null hypothesis (Figure 11). Lastly, the KPSS test affirms the nonstationary nature of the series (Figure 12).

DIFFERENCING

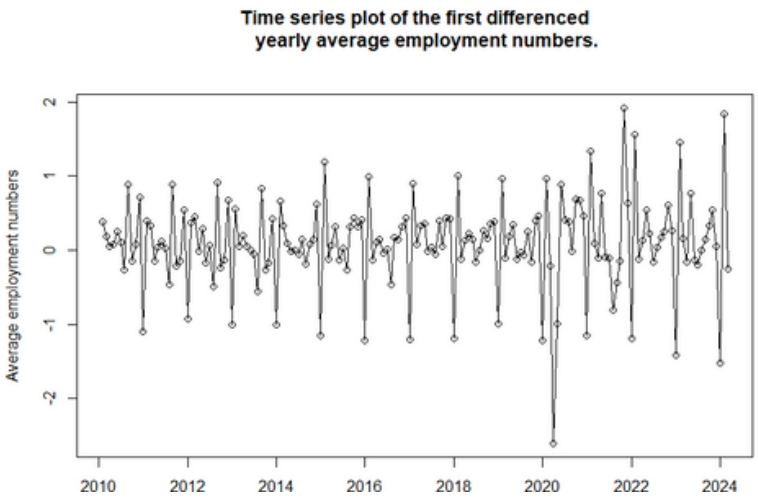

Figure 13: Differenced Time Series

Since the series is non-stationary, differencing was utilized to stabilize the resulting time series around the mean (Figure 13). However, a notable decrease occurred during the COVID period. The ACF and PACF plots post-differencing, illustrated in Figure 14, exhibit a more erratic and less patterned trend. However, it still exhibits a slowly decaying pattern indicating seasonal trend. Furthermore, the stationarity tests depicted in Figures 15-17 validate that the series is stationary, except for the KPSS test, which remained at 0.1. Given the confirmation of stationarity by the other two tests and the more random patterns in the PACF and ACF plots, we opted to use the time series with just one differencing.

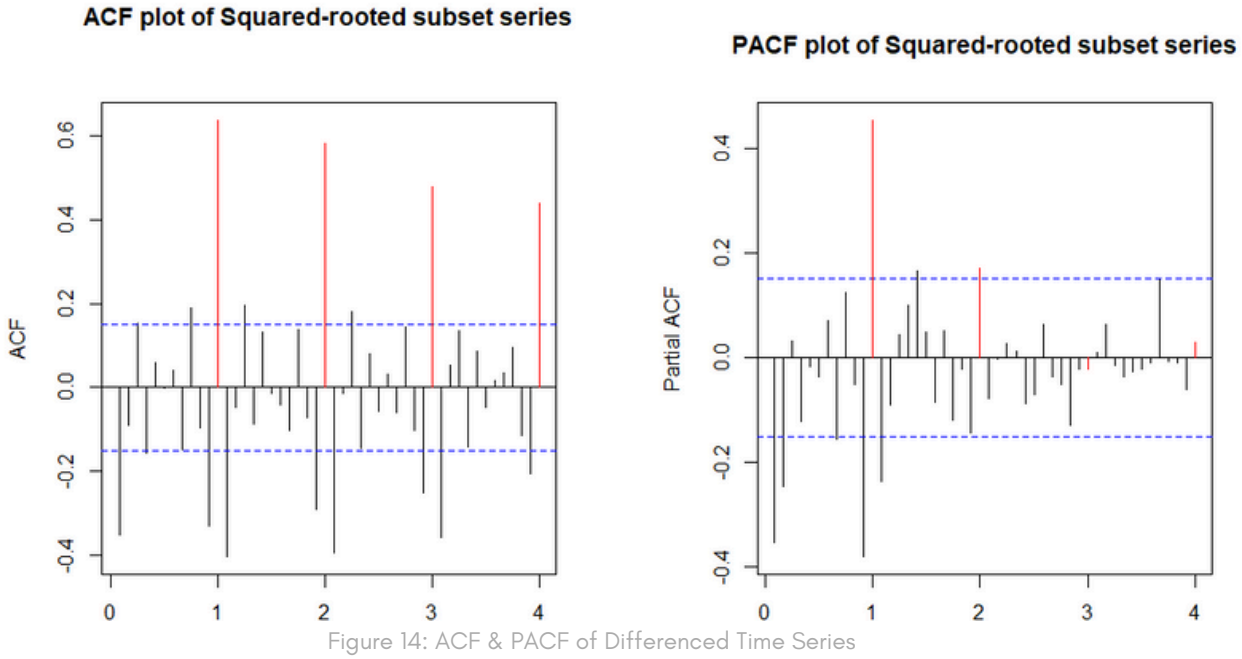

Figure 14: ACF & PACF of Differenced Time Series

```
> adf.test(diff.employment)

    Augmented Dickey-Fuller Test

data:  diff.employment
Dickey-Fuller = -6.4729, Lag order = 5, p-value =
0.01
alternative hypothesis: stationary
```

Figure 15: ADF Test of Employment Time Series

```
> pp.test(diff.employment)

    Phillips-Perron Unit Root Test

data:  diff.employment
Dickey-Fuller Z(alpha) = -202.64, Truncation lag
parameter = 4, p-value = 0.01
alternative hypothesis: stationary
```

Figure 16: PP Test of Employment Time Series

```
> kpss.test(diff.employment)

    KPSS Test for Level Stationarity

data:  diff.employment
KPSS Level = 0.10424, Truncation lag parameter = 4,
p-value = 0.1
```

Figure 17: KPS Test of Employment Time Series

SEASONAL ARIMA

Model 1: D = 1

The ACF and PACF of Model 1 is shown in Figure 18. The seasonal trend in the series has successfully been removed from the series. However, there is a significant autocorrelation at the second seasonal lag in the ACF plot, hence we set the value of Q=1 and P=0 for the next model as there is no significant seasonal lag in the PACF.

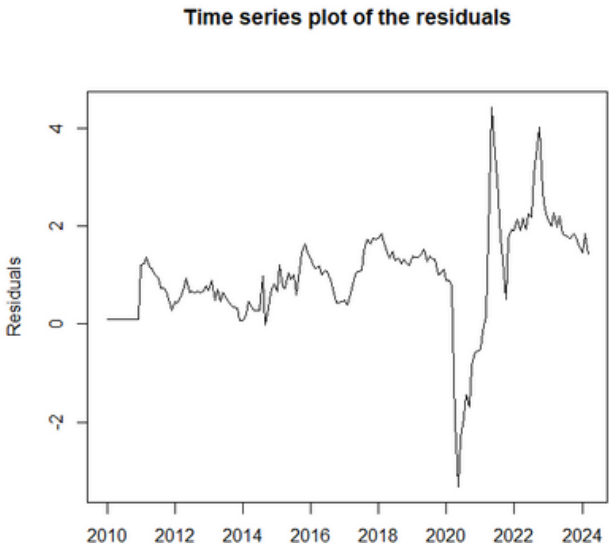

Figure 18: Time Series Plot of the Residuals - m1

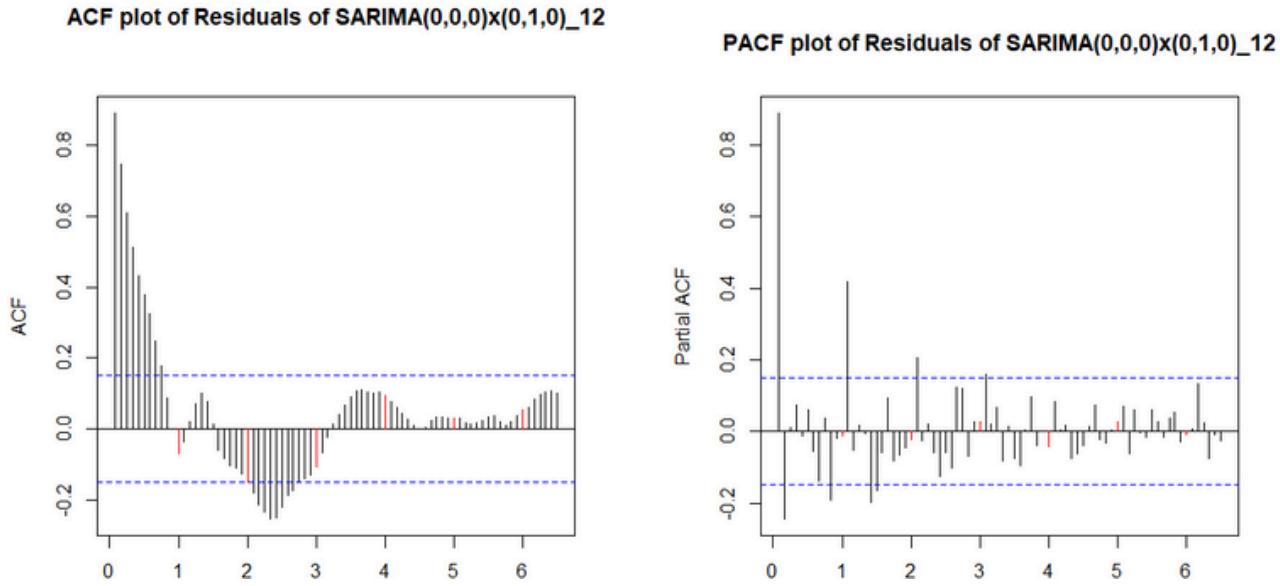

Figure 19: ACF & PACF - m1

Model 2: SARIMA(0,0,0)x(0,1,1)

The initial seasonal lag in the ACF is crucial, as shown in Figure 21. However, the subsequent lags do not show significance, and there is no clear seasonal pattern in the plot (Figure 20). Therefore, we continue to modify this model.

We have now determined the seasonal orders of the model. Accordingly, we will focus on establishing the regular orders of the model while maintaining a consistent seasonal order.

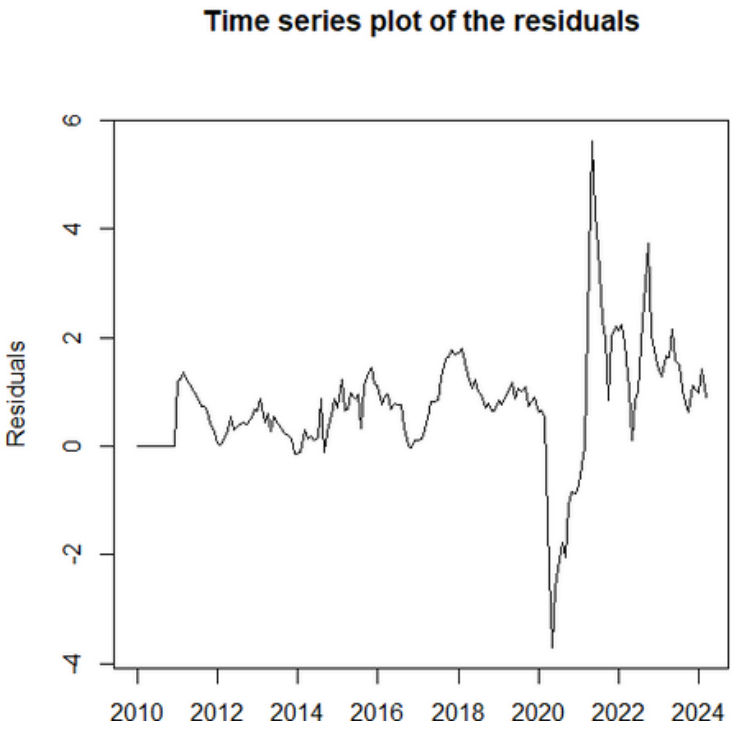

Figure 20: Time Series Plot of the Residuals - m2

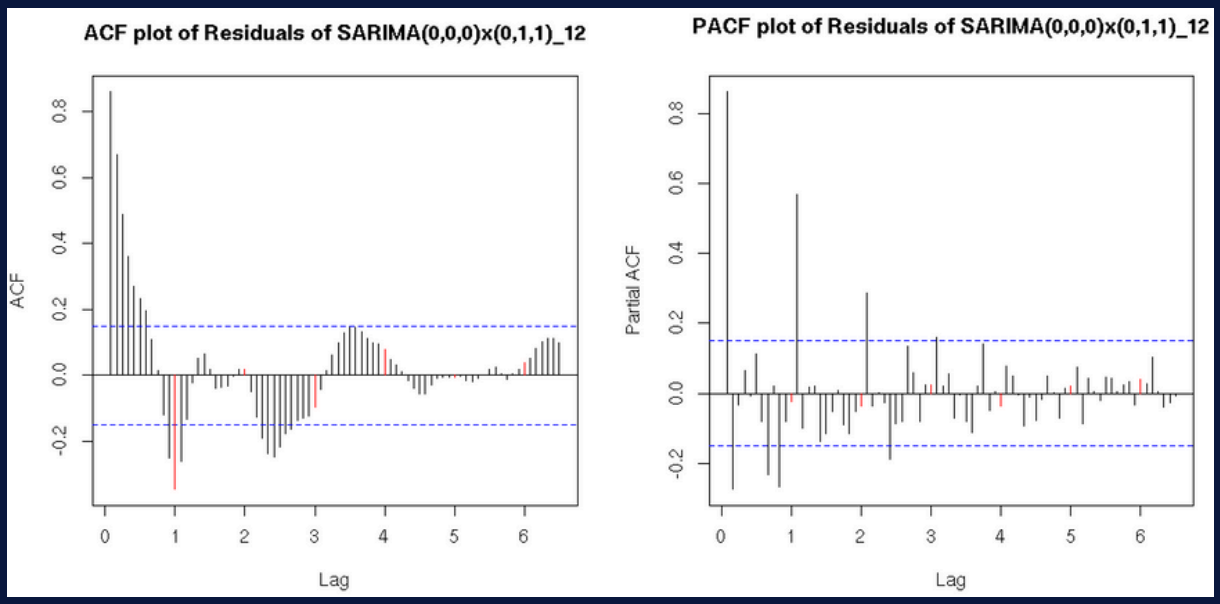

Figure 21: ACF & PACF - m2

Model 3: SARIMA(0,1,0)x(0,1,1)

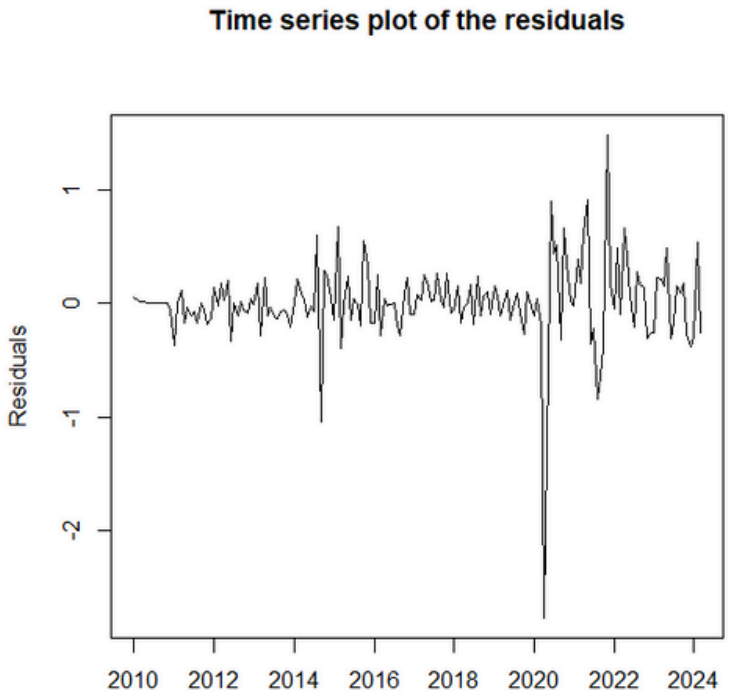

Figure 22: Time Series Plot of the Residuals - m3

Initially, we utilize ordinary differencing on the series to render it stationary, following previous experiments. By examining Model 3, we identify the first potential set of ordinary orders by observing notable lags between lag 0 and lag 1 in the ACF and PACF plots. With 3 lags in the PACF and 2 in the ACF, we establish p=3 and q=2.

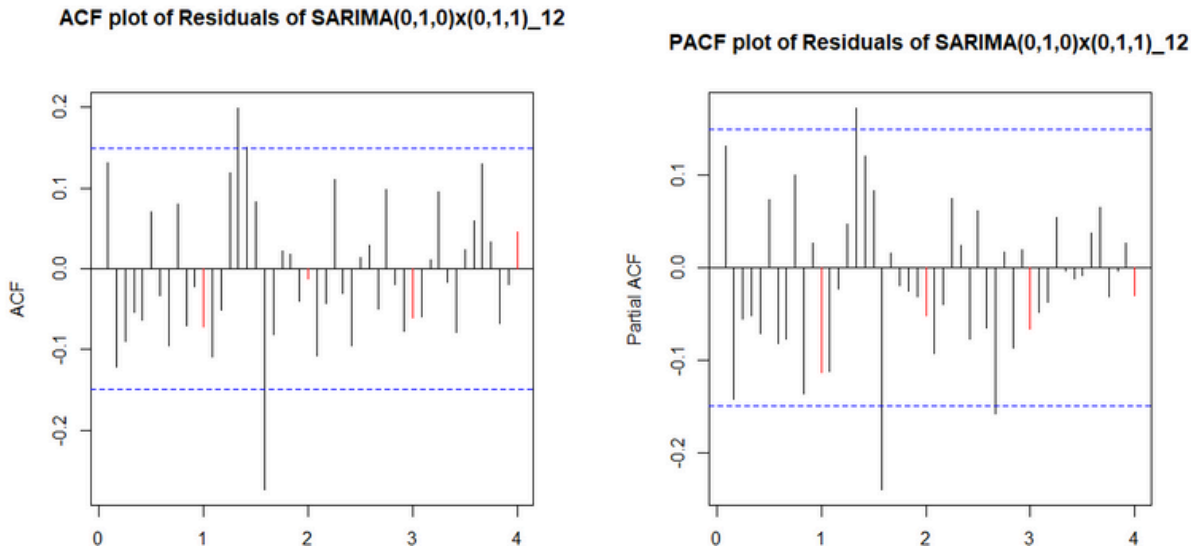

Figure 23: ACF & PACF - m3

EACF

To identify alternative sets of parameters for our SARIMA model, we follow the same process used for determining the orders in an ARIMA model.

As seen in Figure 26, the top left model on the EACF that is not distracted by any x is at (0,2). We chose (0,2) instead of (0,1) as our previous analysis seems to be pointing to larger models to capture the complexity of the data. (0,2) will also allow us to explore more neighbor models than (0,1). We also consider its neighbours: (0,3) (1,2) (1,3). Our list of possible models now include:

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	o	o	o
1	x	x	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	o	o	o	o	o	o	o	o	o	o	o	o
4	x	o	x	o	o	o	o	o	o	o	o	o	o	o
5	x	o	x	o	x	o	o	o	o	o	o	o	o	o
6	x	o	x	o	x	o	o	o	o	o	o	o	o	o
7	x	x	x	o	o	o	o	o	o	o	o	o	o	o

Figure 26: EACF of Res 3

- SARIMA(3,1,2)x(0,1,1)_12
- SARIMA(0,1,2)x(0,1,1)_12
- SARIMA(0,1,3)x(0,1,1)_12
- SARIMA(1,1,2)x(0,1,1)_12
- SARIMA(1,1,3)x(0,1,1)_12

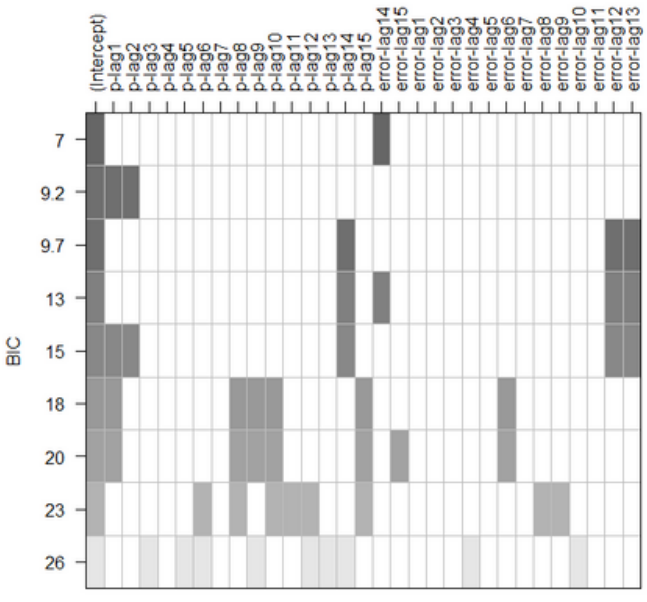

Figure 27: BIC table of Res 3

BIC

BIC table: Next, we look at the BIC table to find more possible models. Since bigger models are usually able to capture more trend and information, we increase the limits in the BIC table to 15 to see if there are any potential big models for our series. Figure 27 shows that the best model in the BIC table has a MA order of 14 and an AR order of 0. This is a big model and we add it to our list of models to consider.

- SARIMA(3,1,2)x(0,1,1)_12
- SARIMA(0,1,2)x(0,1,1)_12
- SARIMA(0,1,3)x(0,1,1)_12
- SARIMA(1,1,2)x(0,1,1)_12
- SARIMA(1,1,3)x(0,1,1)_12
- SARIMA(0,1,14)x(0,1,1)_12

This concludes the final phase of our model specification stage. We have a total of 6 potential models for our employment series. To identify the most suitable model, we will fit each model to the employment series, analyze their residuals, and compare each model using a range of error metrics.

MODEL FITTING & ANALYSIS

SARIMA(3,1,2)x(0,1,1)_12

As seen in Figure 28, both the CSS and ML models found SMA1 to be significant. However, none of the ARIMA variables were found to be significant using the CSS model. In contrast, for the ML model, one ARIMA variable, specifically AR1, was significant. In Figure 29, which shows the residual analysis for the model, the time series plot indicated a clear change point and changing variance over time, while the histogram and QQ plot showed that the residuals were not normally distributed but were left-skewed. The Shapiro-Wilk test also showed the same thing with p-value = 5.49e-15. Lastly, the ACF and Ljung-Box test results indicated that there were still significant autocorrelations present in the residuals.

```
> m312 = fit_SARIMA(sqrt_data,orders=c(3,1,2))
[1] "CSS Method"

z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ar1   0.578262   0.486527   1.1886   0.2346
ar2   0.234867   0.648342   0.3623   0.7172
ar3  -0.030418   0.166715  -0.1825   0.8552
ma1  -0.460402   0.475089  -0.9691   0.3325
ma2  -0.431747   0.534029  -0.8085   0.4188
sma1 -0.683823   0.061125 -11.1873   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[1] "ML Method"

z test of coefficients:

      Estimate Std. Error z value Pr(>|z|)
ar1   1.476431   0.661461  2.2321  0.02561 *
ar2  -0.686678   0.696519 -0.9859  0.32420
ar3   0.162038   0.124098  1.3057  0.19165
ma1  -1.346803   0.666310 -2.0213  0.04325 *
ma2   0.363634   0.641142  0.5672  0.57060
sma1 -0.644415   0.066049 -9.7566  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 28: Coeftest - SARIMA(3,1,2)x(0,1,1)_12

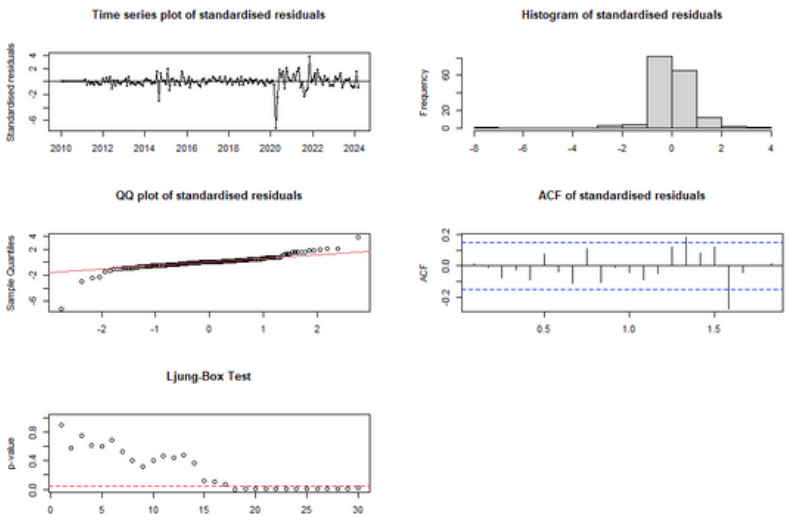

Figure 29: Residual Analysis- SARIMA(3,1,2)x(0,1,1)_12

```
> m012 = fit_SARIMA(sqrt_data,orders=c(0,1,2))
[1] "CSS Method"

z test of coefficients:

     Estimate Std. Error  z value Pr(>|z|)
ma1   0.125418   0.083774   1.4971   0.1344
ma2  -0.120258   0.086305  -1.3934   0.1635
sma1 -0.709530   0.055833 -12.7080   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] "ML Method"

z test of coefficients:

     Estimate Std. Error  z value Pr(>|z|)
ma1   0.132907   0.084258   1.5774   0.1147
ma2  -0.106295   0.086470  -1.2293   0.2190
sma1 -0.676392   0.063993 -10.5698   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 30: Coeftest - SARIMA(0,1,2)x(0,1,1)_12

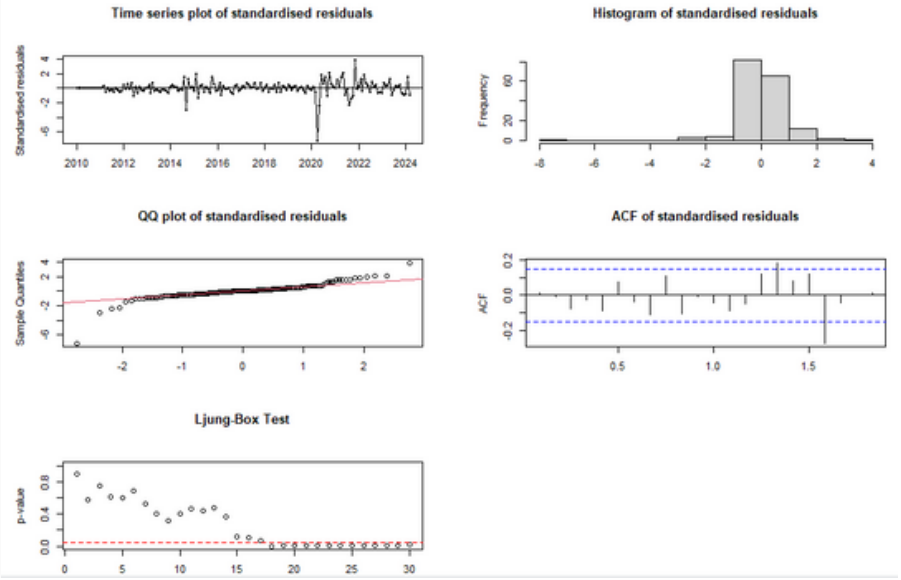

Figure 31: Residual Analysis - SARIMA(0,1,2)x(0,1,1)_12

SARIMA(0,1,2)x(0,1,1)_12

In the coefficient test for both the CSS and ML models, while the seasonable variable SMA1 was signficant for both models, no ARIMA variables were found to be significant (Figure 30). The residual analysis revealed several key observations (Figure 31). The time series plot indicated a clear change point and changing variance over time. The histogram and QQ plot showed that the residuals were not normally distributed and exhibited a left-skew. The Shapiro-Wilk test also showed the same thing with p-value = 1.852e-14. Furthermore, the ACF and Ljung-Box test results indicated that there were still significant autocorrelations present in the residuals.

SARIMA(0,1,3)x(0,1,1)_12

Upon analyzing the CSS and ML models, no noteworthy ARIMA variable was identified, though the seasonal variable SMA1 exhibited significance (Figure 32) . Notable observations from the analysis include a discernible change point and fluctuating variance in the time series plot. Both the histogram and QQ plot suggest a left-skewed distribution of residuals, deviating from normality. The Shapiro-Wilk test also showed the same thing with p-value = 1.94e-14. Furthermore, the ACF and Ljung-box test indicate the presence of sustained significant autocorrelations within the residual data (Figure 33).

```
> m013 = fit_SARIMA(sqrt_data,orders=c(0,1,3))
[1] "CSS Method"

z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ma1   0.137639   0.081889   1.6808  0.09280 .
ma2  -0.146587   0.085446  -1.7155  0.08624 .
ma3  -0.098985   0.084001  -1.1784  0.23865
sma1 -0.693698   0.058397 -11.8789  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[1] "ML Method"

z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ma1   0.143427   0.081478   1.7603  0.07835 .
ma2  -0.136474   0.084848  -1.6084  0.10774
ma3  -0.109645   0.082460  -1.3297  0.18363
sma1 -0.659534   0.065567 -10.0589  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 32: Coeftest - SARIMA(0,1,3)x(0,1,1)_12

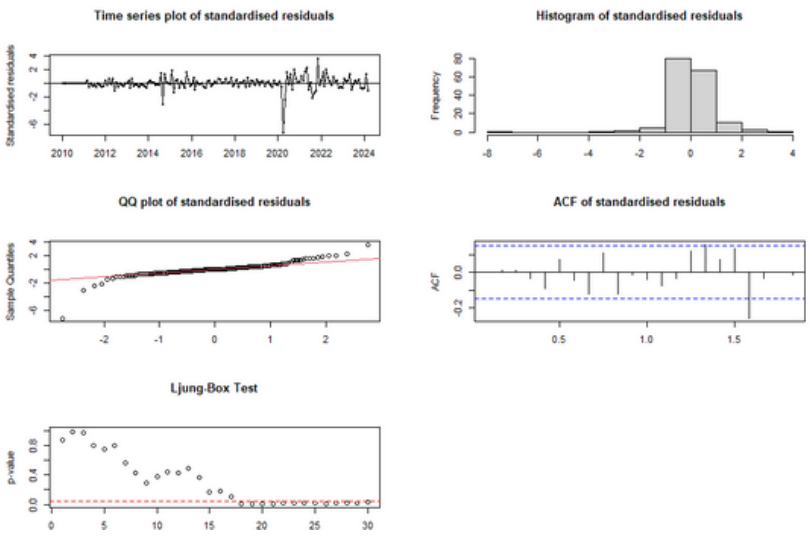

Figure 33: Residual Analysis - SARIMA(0,1,3)x(0,1,1)_12

SARIMA(1,1,2)x(0,1,1)_12

In the coefficient test for the CSS model, one ARIMA variable (MA2) was found to be significant. In contrast, for the ML model, all ARIMA variables were significant. Additionally, the seasonal variable SMA1 was significant in both models (Figure 34). The residual analysis revealed several important observations. The time series plot indicated a clear change point and changing variance over time. The histogram and QQ plot demonstrated that the residuals were not normally distributed and were left-skewed. The Shapiro-Wilk test also showed the same thing with p-value = 1.098e-14. Moreover, the ACF and Ljung-Box test results showed that there were still significant autocorrelations present in the residuals (Figure 35).

```
> m112 = fit_SARIMA(sqrt_data,orders=c(1,1,2))
[1] "CSS Method"

z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1   0.745066   0.380730   1.9569  0.050354 .
ma1  -0.611997   0.367177  -1.6668  0.095562 .
ma2  -0.222787   0.080159  -2.7793  0.005448 **
sma1 -0.690401   0.060245 -11.4598 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] "ML Method"

z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
ar1   0.854642   0.131733  6.4877 8.717e-11 ***
ma1  -0.717643   0.149027 -4.8155 1.468e-06 ***
ma2  -0.220217   0.082909 -2.6561  0.007904 **
sma1 -0.646697   0.066436 -9.7342 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 34: Coeftest - SARIMA(1,1,2)x(0,1,1)_12

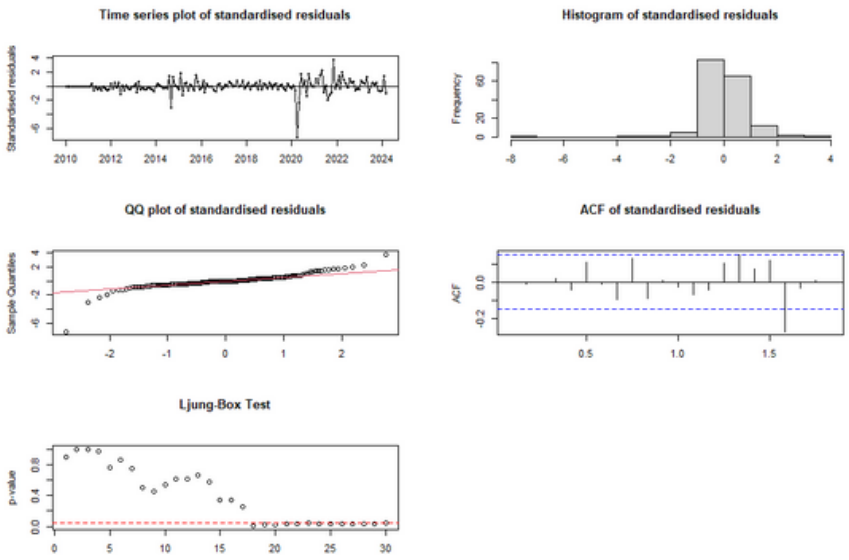

Figure 35: Residual Analysis - SARIMA(1,1,2)x(0,1,1)_12

SARIMA(1,1,3)x(0,1,1)_12

In Figure 36, the coefficient test for both the CSS and ML model showed that all ARIMA variables were found to be significant except for MA3 variable. Additionally, the seasonal variable SMA1 was significant. In the residual analysis, the time series plot indicated a clear change point and changing variance over time. The histogram and QQ plot showed that the residuals were not normally distributed and exhibited a left-skew. The Shapiro-Wilk test also showed the same thing with p-value = 1.431e-14. Moreover, the ACF and Ljung-Box test results indicated that there were still significant autocorrelations present in the residuals (Figure 37).

```
> m113 = fit_SARIMA(sqrt_data,orders=c(1,1,3))
[1] "CSS Method"

z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1   0.928341   0.059963  15.4820 < 2.2e-16 ***
ma1  -0.802056   0.105656  -7.5912 3.169e-14 ***
ma2  -0.243404   0.094634  -2.5721   0.01011 *
ma3   0.059587   0.094229   0.6324   0.52715
sma1 -0.684501   0.061628 -11.1070 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[1] "ML Method"

z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
ar1   0.877968   0.108915  8.0611 7.564e-16 ***
ma1  -0.749510   0.145310 -5.1580 2.496e-07 ***
ma2  -0.233721   0.092496 -2.5268   0.01151 *
ma3   0.031638   0.100778  0.3139   0.75357
sma1 -0.651791   0.067476 -9.6596 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 36: Coeftest - SARIMA(1,1,2)x(0,1,1)_12

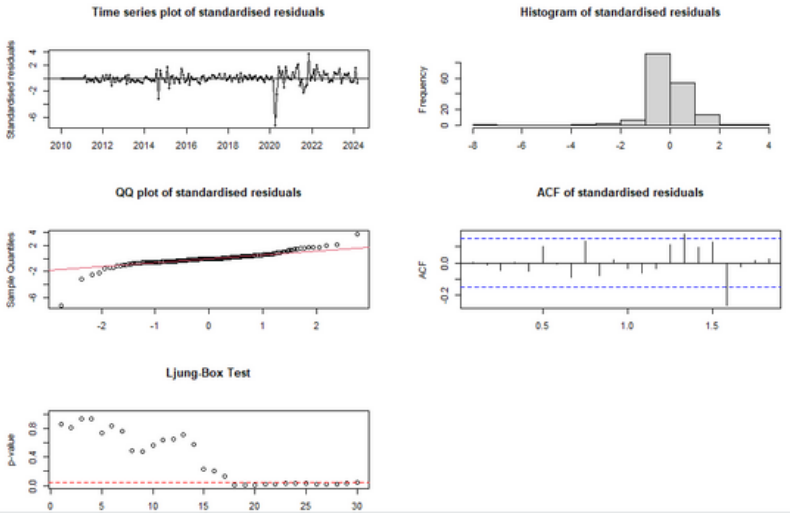

Figure 37: Residual Analysis -SARIMA(1,1,2)x(0,1,1)_12

SARIMA(0,1,14)x(0,1,1)_12

In the coefficient test for the CSS model, four ARIMA variables (MA7, MA9, MA11, and MA12) were found to be significant. For the ML model, only one ARIMA variable (MA12) was significant (Figure 38). The seasonal variable SMA1 was found to be insignificant. The residual analysis revealed several important observations. The time series plot indicated a clear change point and changing variance over time. The histogram and QQ plot showed that the residuals were not normally distributed and exhibited a left-skew. The Shapiro-Wilk test also showed the same thing with p-value = 1.201e-14. The ACF analysis indicated that there were still two significant autocorrelations, while the Ljung-Box test showed no significant autocorrelations (Figure 39).

```
> m0114 = fit_SARIMA(sqrt_data,orders=c(0,1,14))
[1] "CSS Method"

z test of coefficients:

       Estimate Std. Error  z value Pr(>|z|)
ma1   0.1330704  0.0861793   1.5441  0.12256
ma2  -0.1012799  0.0912192  -1.1103  0.26687
ma3   0.0439129  0.0594577   0.7386  0.46018
ma4   0.0809361  0.0602326   1.3437  0.17904
ma5  -0.0898153  0.0612211  -1.4671  0.14236
ma6  -0.0073011  0.0536156  -0.1362  0.89168
ma7  -0.1439501  0.0729206  -1.9741  0.04837 *
ma8  -0.0191146  0.0638677  -0.2993  0.76472
ma9   0.1486580  0.0605349   2.4557  0.01406 *
ma10 -0.0378077  0.0608685  -0.6211  0.53451
ma11 -0.1555092  0.0693069  -2.2438  0.02485 *
ma12 -0.7701234  0.0726501 -10.6004  < 2e-16 ***
ma13 -0.1306505  0.1085987  -1.2031  0.22895
ma14  0.0107825  0.1038928   0.1038  0.91734
sma1 -0.1313602  0.1211771  -1.0840  0.27835
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] "ML Method"

z test of coefficients:

       Estimate Std. Error z value  Pr(>|z|)
ma1   0.1567837  0.1129443  1.3882    0.1651
ma2  -0.1267410  0.1023288 -1.2386    0.2155
ma3   0.0044503  0.0854357  0.0521    0.9585
ma4   0.1130160  0.0848352  1.3322    0.1828
ma5  -0.0496959  0.0994658 -0.4996    0.6173
ma6   0.0798614  0.1109128  0.7200    0.4715
ma7  -0.1181341  0.1079000 -1.0948    0.2736
ma8  -0.0351972  0.0950792 -0.3702    0.7112
ma9   0.1009818  0.0901793  1.1198    0.2628
ma10 -0.0691711  0.0809020 -0.8550    0.3926
ma11 -0.1288922  0.1077076 -1.1967    0.2314
ma12 -0.7417838  0.1426723 -5.1992 2.001e-07 ***
ma13 -0.1705568  0.1036899 -1.6449    0.1000 .
ma14 -0.0148465  0.1052026 -0.1411    0.8878
sma1 -0.1134188  0.1309759 -0.8660    0.3865
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 38: Coeftest - SARIMA(0,1,14)x(0,1,1)_12

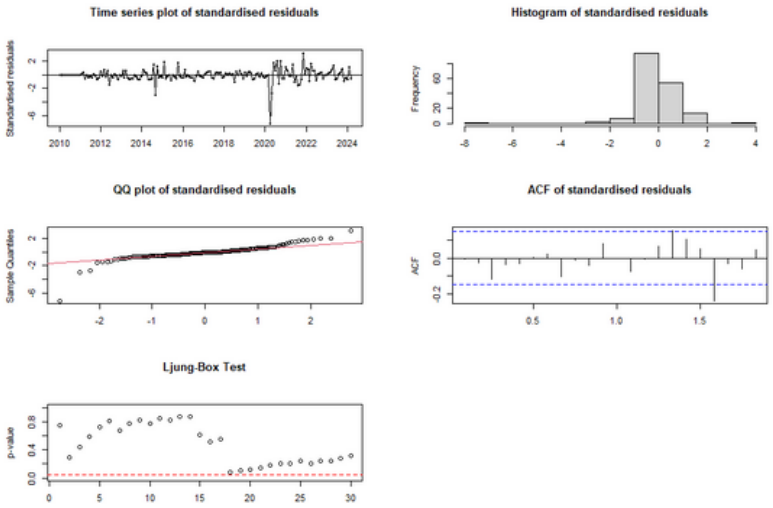

Figure 39: Residual Analysis -SARIMA(0,1,14)x(0,1,1)_12

Accordingly, based on the residual analysis, SARIMA(0,1,14)x(0,1,1)_12 stands out as the optimal model for capturing a greater amount of information compared to the others. This outcome aligns with our expectations since it is a comprehensive model. Notably, all models exhibit a significant outlier in their residuals attributed to the impact of COVID-19.

```
> sort.score(sc.AIC, score = "aic")
          df      AIC
m112$ml    5 155.9362
m012$ml    4 156.6565
m013$ml    5 156.8892
m113$ml    6 157.8402
m312$ml    7 159.2791
m0114$ml  16 163.0327
> sort.score(sc.BIC, score = "bic")
          df      BIC
m012$ml    4 168.9069
m112$ml    5 171.2491
m013$ml    5 172.2022
m113$ml    6 176.2158
m312$ml    7 180.7172
m0114$ml  16 212.0342
```

Figure 40: AIC & BIC Rankings

AIC & BIC Scores

SARIMA(0,1,14)x(0,1,1)_12 holds the lowest rank in both tables. As the difference in AIC scores is quite minor, we still view it as a possible model and will utilize error metrics for further model evaluation.

Error Metrics

The error metric table, as seen in Figure 41, indicates that although SARIMA(0,1,14)x(0,1,1)_12 is a complex model, it is not overfitting the data since its errors metrics are similar to the smaller models. Additionally, SARIMA(0,1,14)x(0,1,1)_12 achieves the highest error scores in 4 out of 7 metrics, solidifying its status as the most promising model.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
SARIMA(3,1,2)x(0,1,1)_12	0.031	0.361	0.208	0.026	0.186	0.169	-0.003
SARIMA(0,1,2)x(0,1,1)_12	0.004	0.365	0.215	0.003	0.193	0.174	0.010
SARIMA(0,1,3)x(0,1,1)_12	0.005	0.363	0.214	0.003	0.192	0.173	0.001
SARIMA(1,1,2)x(0,1,1)_12	0.006	0.362	0.211	0.004	0.189	0.171	-0.002
SARIMA(1,1,3)x(0,1,1)_12	-0.017	0.362	0.214	-0.018	0.192	0.173	0.001
SARIMA(0,1,14)x(0,1,1)_12	-0.026	0.335	0.200	-0.024	0.180	0.162	-0.004

Figure 41: Error Metric Scores

OVERFITTING MODEL

So far the best model for our series is SARIMA(0,1,14)x(0,1,1)_12. To ensure that we find the most suitable model, we also try 2 more models which are the overfitted models of SARIMA(0,1,14)x(0,1,1)_12.

SARIMA(1,1,14)x(0,1,1)_12

In the coefficient test for the CSS model (Figure 42), five ARIMA variables (MA1, MA9, MA11, MA12, and MA13) were found to be significant. In contrast, the ML model showed no significant variables, including the seasonal variable. In the residual analysis (Figure 43), the time series plot indicated a clear change point and changing variance over time. The histogram and QQ plot demonstrated that the residuals were not normally distributed and were left-skewed. The Shapiro-Wilk test also showed the same thing with p-value = 2.899e-16. The ACF analysis showed that there were still two significant autocorrelations, while the Ljung-Box test indicated no significant autocorrelations.

```
> m1114 = fit_SARIMA(sqrt_data,orders=c(1,1,14)) #About half of the vars are significant
[1] "CSS Method"

z test of coefficients:

       Estimate  Std. Error  z value Pr(>|z|)
ar1  -0.56246551  0.32514750  -1.7299  0.08365 .
ma1   0.72211364  0.32510799   2.2212  0.02634 *
ma2   0.04239578  0.11308344   0.3749  0.70773
ma3   0.00030998  0.07488615   0.0041  0.99670
ma4   0.10644534  0.06804706   1.5643  0.11775
ma5  -0.05791418  0.06463131  -0.8961  0.37022
ma6  -0.05744363  0.06472432  -0.8875  0.37480
ma7  -0.12924728  0.07164556  -1.8040  0.07123 .
ma8  -0.08860017  0.08502004  -1.0421  0.29736
ma9   0.13961570  0.06893570   2.0253  0.04284 *
ma10  0.04282763  0.07992058   0.5359  0.59204
ma11 -0.16589662  0.06869614  -2.4149  0.01574 *
ma12 -0.85322005  0.08395184 -10.1632  < 2e-16 ***
ma13 -0.60554692  0.26590340  -2.2773  0.02277 *
ma14 -0.16563783  0.12346693  -1.3416  0.17974
sma1 -0.13456763  0.12043853  -1.1173  0.26386
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] "ML Method"

z test of coefficients:

      Estimate Std. Error z value Pr(>|z|)
ar1  -0.5476264  0.3317463 -1.6507  0.09879 .
ma1   0.7178303  0.4694328  1.5291  0.12623
ma2  -0.0021381  0.4457770 -0.0048  0.99617
ma3  -0.0688325  0.1698111 -0.4053  0.68522
ma4   0.1188330  0.2489861  0.4773  0.63317
ma5   0.0140143  0.4692269  0.0299  0.97617
ma6   0.0711681  0.6138387  0.1159  0.90770
ma7  -0.0493889  0.5824500 -0.0848  0.93242
ma8  -0.0912178  0.4422906 -0.2062  0.83660
ma9   0.0752080  0.2123603  0.3542  0.72322
ma10 -0.0262379  0.2067723 -0.1269  0.89903
ma11 -0.1647926  0.4655999 -0.3539  0.72339
ma12 -0.8118068  0.5865573 -1.3840  0.16635
ma13 -0.6024896  0.4143459 -1.4541  0.14593
ma14 -0.1799263  0.1360235 -1.3228  0.18592
sma1 -0.1193744  0.1218009 -0.9801  0.32705
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 42: Coeftest - SARIMA(1,1,14)x(0,1,1)_12

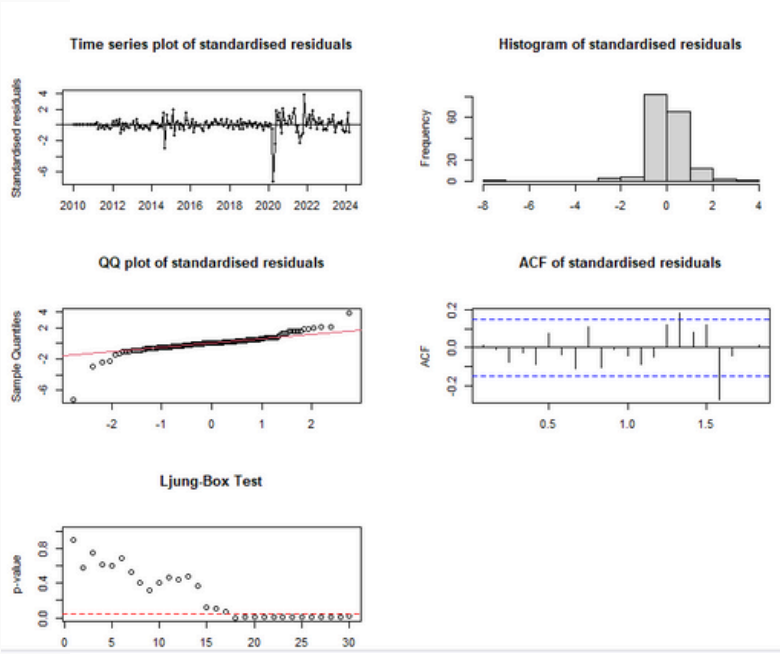

Figure 43: Residual Analysis -SARIMA(1,1,14)x(0,1,1)_12

```
> m0115 = fit_SARIMA(sqrt_data,orders=c(0,1,15)) #Most vars are significant in CSS
[1] "CSS Method"

z test of coefficients:

      Estimate Std. Error z value Pr(>|z|)
ma1   0.135859   0.078910  1.7217  0.08513 .
ma2  -0.122049   0.077310 -1.5787  0.11440
ma3  -0.084396   0.081959 -1.0297  0.30313
ma4   0.066080   0.056208  1.1756  0.23974
ma5  -0.127616   0.063755 -2.0017  0.04532 *
ma6  -0.013170   0.057740 -0.2281  0.81957
ma7  -0.163537   0.078016 -2.0962  0.03606 *
ma8  -0.035225   0.069816 -0.5045  0.61388
ma9   0.115862   0.070932  1.6334  0.10238
ma10 -0.060118   0.058013 -1.0363  0.30007
ma11 -0.113260   0.073794 -1.5348  0.12483
ma12 -0.727551   0.079145 -9.1926  < 2e-16 ***
ma13 -0.149747   0.098012 -1.5278  0.12655
ma14  0.042680   0.088161  0.4841  0.62831
ma15  0.226527   0.096115  2.3568  0.01843 *
sma1 -0.181871   0.130467 -1.3940  0.16332
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] "ML Method"

z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
ma1   0.146565   0.089808  1.6320   0.10268
ma2  -0.153078   0.085618 -1.7879   0.07379 .
ma3  -0.102405   0.102791 -0.9962   0.31913
ma4   0.106194   0.073030  1.4541   0.14592
ma5  -0.084032   0.082549 -1.0180   0.30869
ma6   0.060281   0.093636  0.6438   0.51972
ma7  -0.131118   0.089602 -1.4633   0.14338
ma8  -0.029544   0.076363 -0.3869   0.69884
ma9   0.101606   0.091141  1.1148   0.26493
ma10 -0.065907   0.069927 -0.9425   0.34593
ma11 -0.080373   0.086839 -0.9255   0.35468
ma12 -0.724183   0.104888 -6.9044 5.043e-12 ***
ma13 -0.176750   0.097117 -1.8200   0.06876 .
ma14  0.044305   0.086695  0.5110   0.60932
ma15  0.219825   0.093104  2.3611   0.01822 *
sma1 -0.128827   0.119165 -1.0811   0.27966
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 44: Coeftest - SARIMA(0,1,15)x(0,1,1)_12

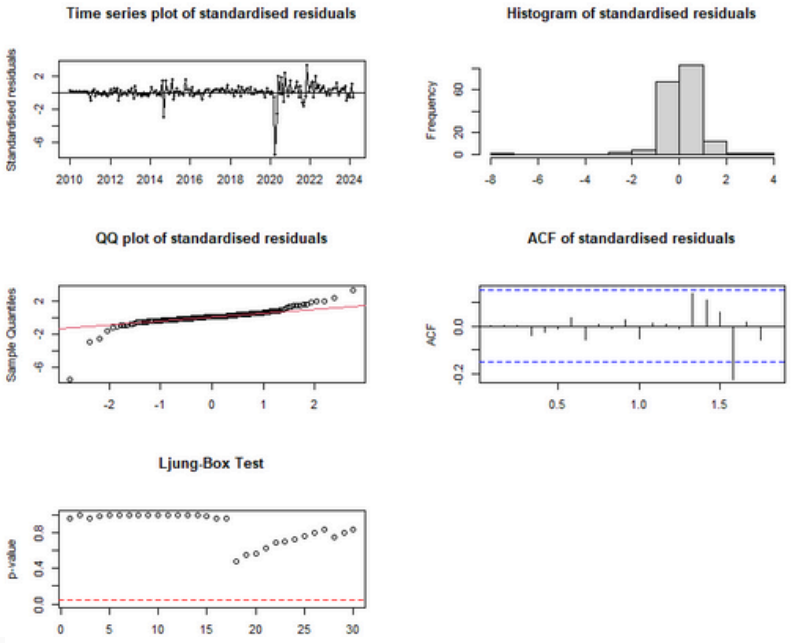

Figure 45: Residual Analysis -SARIMA(0,1,15)x(0,1,1)_12

SARIMA(0,1,15)x(0,1,1)_12

In the coefficient test of the CSS model (Figure 44), four ARIMA variables (MA5, MA7, MA12, and MA15) showed significance. Meanwhile, the ML model highlighted the significance of two ARIMA variables (MA12 and MA15). Notably, the seasonal variable did not prove significant in either model. Upon conducting residual analysis (Figure 45), various key observations emerged. The time series plot pointed out a distinct change point and variance shift over time. Both the histogram and QQ plot revealed non-normally distributed residuals with a left-skewed pattern. The Shapiro-Wilk test also showed the same thing with p-value = 2.2e-16. Additionally, the ACF analysis identified two significant autocorrelations, while the Ljung-Box test indicated no significant autocorrelations.

```
> sort.score(sc.AIC, score = "aic")
          df      AIC
m112$ml    5 155.9362
m012$ml    4 156.6565
m013$ml    5 156.8892
m113$ml    6 157.8402
m312$ml    7 159.2791
m0115$ml  17 159.8167
m0114$ml  16 163.0327
m1114$ml  17 164.1893
> sort.score(sc.BIC, score = "bic")
          df      BIC
m012$ml    4 168.9069
m112$ml    5 171.2491
m013$ml    5 172.2022
m113$ml    6 176.2158
m312$ml    7 180.7172
m0115$ml  17 211.8808
m0114$ml  16 212.0342
m1114$ml  17 216.2534
```

Figure 46: AIC & BIC Scores

AIC & BIC Scores

SARIMA(0,1,15)x(0,1,1)_12 shows improved AIC and BIC scores compared to SARIMA(0,1,14)x(0,1,1)_12, suggesting it may be a more suitable model.

Error Metrics

The error metric table indicates that despite SARIMA(0,1,14)x(0,1,1)_12 being a complex model, it does not exhibit overfitting as its errors align with those of smaller models. Additionally, SARIMA(0,1,14)x(0,1,1)_12 demonstrates the best error scores in 4 out of 7 metrics, positioning it as the most promising model.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
SARIMA(3,1,2)x(0,1,1)_12	0.031	0.361	0.208	0.026	0.186	0.169	-0.003
SARIMA(0,1,2)x(0,1,1)_12	0.004	0.365	0.215	0.003	0.193	0.174	0.010
SARIMA(0,1,3)x(0,1,1)_12	0.005	0.363	0.214	0.003	0.192	0.173	0.001
SARIMA(1,1,2)x(0,1,1)_12	0.006	0.362	0.211	0.004	0.189	0.171	-0.002
SARIMA(1,1,3)x(0,1,1)_12	-0.017	0.362	0.214	-0.018	0.192	0.173	0.001
SARIMA(0,1,14)x(0,1,1)_12	-0.026	0.335	0.200	-0.024	0.180	0.162	-0.004
SARIMA(0,1,15)x(0,1,1)_12	-0.026	0.330	0.195	-0.025	0.175	0.158	-0.010
SARIMA(1,1,14)x(0,1,1)_12	-0.025	0.335	0.198	-0.024	0.178	0.161	-0.021

Figure 47: Error Metric Scores

FINAL JUDGEMENT & FORCASTING

Upon careful analysis and evaluation of various models, SARIMA(0,1,15)x(0,1,1)_12 was selected as the most appropriate model for our data series. The model used the ML method as it produced smaller forecast intervals. The next step involves proceeding to the forecasting stage with this chosen model. As seen in Figures 48 and 49, the next 10-month employment forecast is depicted by 1000. The blue line indicates the point forecast, while the darker shaded areas represent the lower and upper bounds with 80% confidence, and the lighter shaded areas represent the lower and upper bounds with 95% confidence. To further test our model, April's actual employment numbers which was 14,289,600 was compared to our model's predictions (Australian Bureau of Statistics, 2024). This comparison provides a valuable validation of our model's predictive capabilities, indicating that while the model performs well, there may be slight variations due to external factors not accounted for in the model.

```
> forecastML
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Apr 2024       14194.56 14083.67 14305.89 14025.14 14365.00
May 2024       14286.72 14117.92 14456.53 14028.97 14546.82
Jun 2024       14303.34 14101.81 14506.30 13995.71 14614.31
Jul 2024       14264.32 14040.85 14489.56 13923.26 14609.51
Aug 2024       14258.62 14010.78 14508.63 13880.47 14641.86
Sep 2024       14246.56 13979.59 14516.05 13839.29 14659.73
Oct 2024       14285.70 13997.83 14576.49 13846.63 14731.61
Nov 2024       14494.65 14189.70 14802.85 14029.58 14967.31
Dec 2024       14548.16 14229.62 14870.23 14062.42 15042.15
Jan 2025       14230.49 13900.31 14564.55 13727.09 14742.96
```

Figure 48: 10 Month Forecast

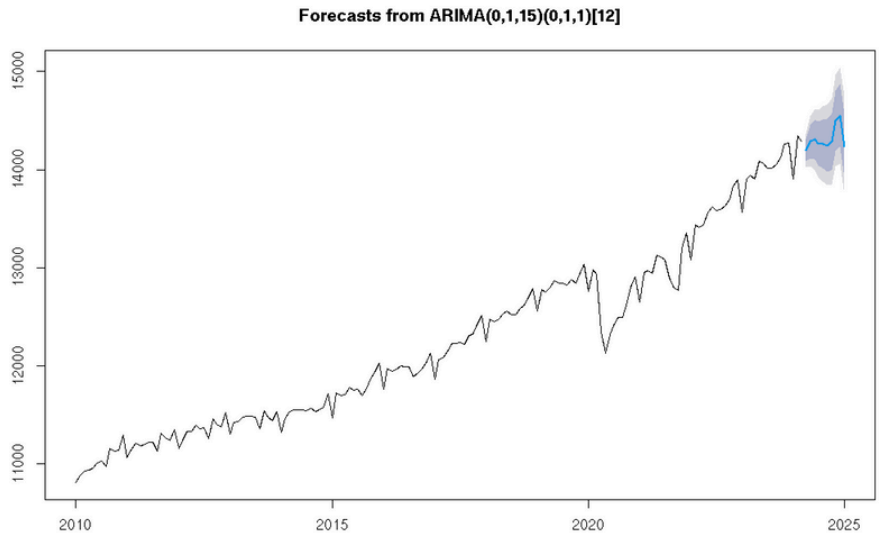

Figure 49: Forecast Graph

CONCLUSION

The SARIMA(0,1,15)x(0,1,1)_12 model has proven to be a suitable choice for forecasting employment numbers in Australia, based on our careful analysis and evaluation. The model's performance, as illustrated by the forecasted employment numbers for the next 10 months, shows a high level of accuracy, with the actual employment rate for April closely aligning with our predictions.

The point forecast and confidence intervals provide a reliable range for expected employment figures, which can be instrumental for policymakers, businesses, and other stakeholders in making informed decisions. Despite the slight discrepancy observed with April's actual employment rate falling just below the 80% confidence interval, the model remains a robust tool for forecasting.

Moving forward, continuous monitoring and refinement of the model will be essential to accommodate any external factors and enhance predictive accuracy. The insights gained from this initial validation will guide further improvements, ensuring that our forecasts remain precise and valuable for supporting economic growth and stability in Australia.

ACKNOWLEDGMENTS

We would like to thank and acknowledge Le Van Tra Tran and Tin Trung Pham who devloped the functions we used to highlight the seasons on ACF and PACF plots for seasonal series. We would also like to thank Haydar Demirhan for his functions we used, particulary the residual analysis function and BIC and AIC functions.

CONTRIBUTION TABLE

No.	**Name of Team Member**	**Contribution to the Project**
1	Faith Ha	.50
2	Chau Le	.50
	Sum:	1

Table 2: Contribution Table

REFERENCES

Australian Bureau of Statistics. (2024). Labour force, Australia. Australian Bureau of Statistics; Australian Bureau of Statistics. https://www.abs.gov.au/statistics/labour/employment-and-unemployment/labour-force-australia/latest-release

CODE APPENDIX

```
rm(list=ls())
library(TSA)
library(tseries)
library(forecast)
library(lmtest)
library(dplyr)

residual.analysis <- function(model, std = TRUE,start = 2, class = c("ARIMA","GARCH","ARMA-GARCH",
"garch", "fGARCH")[1]){
  library(TSA)
  library(FitAR)
  if (class == "ARIMA"){
    if (std == TRUE){
      res.model = rstandard(model)
    }else{
      res.model = residuals(model)
    }
  }else if (class == "GARCH"){
    res.model = model$residuals[start:model$n.used]
  }else if (class == "garch"){
    res.model = model$residuals[start:model$n.used]
  }else if (class == "ARMA-GARCH"){
    res.model = model@fit$residuals
  }else if (class == "fGARCH"){
    res.model = model@residuals
  }else {
    stop("The argument 'class' must be either 'ARIMA' or 'GARCH' ")
  }

  par(mfrow=c(3,2))
  plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of standardised
residuals")
  abline(h=0)
  hist(res.model,main="Histogram of standardised residuals")
  qqnorm(res.model,main="QQ plot of standardised residuals")
  qqline(res.model, col = 2)
  acf(res.model,main="ACF of standardised residuals")
  print(shapiro.test(res.model))
  k=0
  LBQPlot(res.model, lag.max = 30, StartLag = k + 1, k = 0, SquaredQ = FALSE)
  par(mfrow=c(1,1))
}

show_ACF_PACF <- function(series,maxlag,name){
  par(mfrow=c(1,2))
  seasonal_acf(series,lag.max=maxlag, main=paste("ACF plot of",name))
  seasonal_pacf(series, lag.max=maxlag, main=paste("PACF plot of",name))
  par(mfrow=c(1,1))
}


sort.score <- function(x, score = c("bic", "aic")){
  if (score == "aic"){
    x[with(x, order(AIC)),]
  } else if (score == "bic") {
    x[with(x, order(BIC)),]
  } else {
    warning('score = "x" only accepts valid arguments ("aic","bic")')
  }
}
```

```r
# Normalize function
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}


helper <- function(class = c("acf", "pacf"), ...) {

  # Capture additional arguments
  params <- match.call(expand.dots = TRUE)
  params <- as.list(params)[-1]

  # Calculate ACF/PACF values
  if (class == "acf") {
    acf_values <- do.call(acf, c(params, list(plot = FALSE)))
  } else if (class == "pacf") {
    acf_values <- do.call(pacf, c(params, list(plot = FALSE)))
  }

  # Extract values and lags
  acf_data <- data.frame(
    Lag = as.numeric(acf_values$lag),
    ACF = as.numeric(acf_values$acf)
  )

  # Identify seasonal lags to be highlighted
  seasonal_lags <- acf_data$Lag %% 1 == 0

  # Plot ACF/PACF values
  if (class == "acf") {
    do.call(acf, c(params, list(plot = TRUE)))
  } else if (class == "pacf") {
    do.call(pacf, c(params, list(plot = TRUE)))
  }

  # Add colored segments for seasonal lags
  for (i in which(seasonal_lags)) {
    segments(x0 = acf_data$Lag[i], y0 = 0, x1 = acf_data$Lag[i], y1 = acf_data$ACF[i], col = "red")
  }
}


# seasonal_acf ------------------------------------------------------------

seasonal_acf <- function(...) {
  helper(class = "acf", ...)
}


# seasonal_pacf -----------------------------------------------------------

seasonal_pacf <- function(...) {
  helper(class = "pacf", ...)
}

show_ACF_PACF <- function(series,maxlag,name){
  par(mfrow=c(1,2))
```

```r
# read in csv file
employment <- read.csv("employment.csv", header=TRUE)
employment
class(employment)
head(employment)


# Conversion to Time Series
employment_TS <- ts(employment$employed, start = c(1978, 2), frequency = 12)
#Subset the series ot 2010-2024 only
subset = window(employment_TS,2010,)
subset

class(subset)
summary(subset)
x11()
plot(subset,type='l',ylab='Employment', main = " Time series plot of Employment by 1000")
summary(subset)

# acf and pacf plots
show_ACF_PACF(subset,maxlag = 48,name = "the Employment series.")


# adf test (weak indicator of non-stationary)
adf.test(subset, alternative = c("stationary"))
# pp test
pp.test(subset)
# kpss test
kpss.test(subset)

# # MCleod test
# McLeod.Li.test(y=employment_TS,main="McLeod-Li Test Statistics for Monthly Employment Numbers")

# qq plot
qqnorm(subset)
qqline(subset, col = 2)
shapiro.test(subset)

# Normalization
# A small positive constant
log_employment <- log(subset + .001)
min_value <- min(subset)

# qq plot
qqnorm(log_employment)
qqline(log_employment, col = 2)
shapiro.test(log_employment)


normalized_ts <- normalize(subset)

# qq plot
qqnorm(normalized_ts)
qqline(normalized_ts, col = 2)
shapiro.test(normalized_ts)
```

```r
# Box Cox Transformation
BC <- BoxCox.ar(subset, lambda = seq(0.5, 1, 0.01))
BC$ci
lambda <- BC$lambda[which(max(BC$loglike) == BC$loglike)]
lambda # lambda = 0.5 corresponds to square root transformation?

sqrt_data <- sqrt(subset)

plot(sqrt_data,type='l',ylab= 'Employment Numbers', main = 'Time series plot of BC transformed
     monthly average employment numbers.')

# acf and pacf plots
par(mfrow=c(1,2))
show_ACF_PACF(sqrt_data,maxlag = 48,name = "the Employment series.")
par(mfrow=c(1,1))

# differencing

diff.employment = diff(sqrt_data)

par(mfrow=c(1,2))
show_ACF_PACF(diff.employment,maxlag=48,name="Squared-rooted subset series")
par(mfrow=c(1,1))

# adf, pp, kpss test
adf.test(diff.employment)
pp.test(diff.employment)
kpss.test(diff.employment)

#SEASONAL ARIMA
sqrt_data
#Start with a baseline model with D=1
m1 = Arima(sqrt_data,order=c(0,0,0),seasonal=list(order=c(0,1,0), period=12))
res.m1 = residuals(m1)
par(mfrow=c(1,1))
plot(res.m1,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
show_ACF_PACF(res.m1, maxlag = 78, name="Residuals of SARIMA(0,0,0)x(0,1,0)_12")

#From the ACF and PACF => P=0 and Q=2
m2 = Arima(sqrt_data,order=c(0,0,0),seasonal=list(order=c(0,1,1), period=12),method="CSS")
res.m2 = residuals(m2)
par(mfrow=c(1,1))
plot(res.m2,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
show_ACF_PACF(res.m2, maxlag = 78,name="Residuals of SARIMA(0,0,0)x(0,1,1)_12")
adf.test(res.m2) #Residuals are stationary

#Add differencing
m3 = Arima(sqrt_data,order=c(0,1,0),seasonal=list(order=c(0,1,1), period=12))
res.m3 = residuals(m3);
par(mfrow=c(1,1))
plot(res.m3,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
show_ACF_PACF(res.m3, maxlag = 48,name="Residuals of SARIMA(0,1,0)x(0,1,1)_12")
#No trend or seasonality left, move forward
adf.test(res.m3)
```

```
#From ACF and PACF of m3
#p = 3 and q = 2
#Determined by significant bars between 0 and 1
m4 = Arima(sqrt_data,order=c(3,1,2),
           seasonal=list(order=c(0,1,1), period=12))
res.m4 = residuals(m4);
par(mfrow=c(1,1))
plot(res.m4,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
show_ACF_PACF(res.m4, maxlag = 48,name="Residuals of SARIMA(3,1,3)x(0,1,2)_12")
# SARIMA(3,1,2)x(0,1,1)_12

eacf(res.m3)
# The tentative models are specified as
# SARIMA(0,1,2)x(0,1,1)_12
# SARIMA(0,1,3)x(0,1,1)_12
# SARIMA(1,1,2)x(0,1,1)_12
# SARIMA(1,1,3)x(0,1,1)_12

par(mfrow=c(1,1))
bic_table = armasubsets(y=res.m3,nar=15,nma=15,y.name='p',ar.method='ols')
plot(bic_table)
# SARIMA(0,1,14)x(0,1,1)_12


#Final sets of possible models
# SARIMA(3,1,2)x(0,1,1)_12
# SARIMA(0,1,2)x(0,1,1)_12
# SARIMA(0,1,3)x(0,1,1)_12
# SARIMA(1,1,2)x(0,1,1)_12
# SARIMA(1,1,3)x(0,1,1)_12
# SARIMA(0,1,14)x(0,1,1)_12

#Define a function to fit models quickly
fit_SARIMA <- function(series,orders) {
  model_css = Arima(series,order=orders,
                    seasonal=list(order=c(0,1,1),period=12),
                    method='CSS')
  coef_css = coeftest(model_css)
  print("CSS Method")
  print(coef_css)

  model_ml = Arima(series,order=orders,
                   seasonal=list(order=c(0,1,1),period=12),
                   method='ML')
  coef_ml = coeftest(model_ml)
  print("ML Method")
  print(coef_ml)
  return(list(css=model_css,ml=model_ml))
}
# SARIMA(3,1,2)x(0,1,1)_12
m312 = fit_SARIMA(sqrt_data,orders=c(3,1,2))
residual.analysis(model = m312$css) #Did not capture the info well
# SARIMA(0,1,2)x(0,1,1)_12
m012 = fit_SARIMA(sqrt_data,orders=c(0,1,2))
residual.analysis(model = m012$css) #Did not capture the info well
```

```
# SARIMA(0,1,3)x(0,1,1)_12
m013 = fit_SARIMA(sqrt_data,orders=c(0,1,3))
residual.analysis(model = m013$css) #Did not capture the info well
# SARIMA(1,1,2)x(0,1,1)_12
m112 = fit_SARIMA(sqrt_data,orders=c(1,1,2))
residual.analysis(model = m112$css) #Did not capture the info well
# SARIMA(1,1,3)x(0,1,1)_12
m113 = fit_SARIMA(sqrt_data,orders=c(1,1,3))
residual.analysis(model = m113$css) #Did not capture the info well
# SARIMA(0,1,14)x(0,1,1)_12
m0114 = fit_SARIMA(sqrt_data,orders=c(0,1,14))
residual.analysis(model = m0114$css) #Captured info pretty well
sc.AIC = AIC(m312$ml,m012$ml,m013$ml,m112$ml,m113$ml,m0114$ml)
sc.BIC = BIC(m312$ml,m012$ml,m013$ml,m112$ml,m113$ml,m0114$ml)
sort.score(sc.AIC, score = "aic")
sort.score(sc.BIC, score = "bic")
#The big model captured information well but have the worst aic and bic scores

# error metrics for first 6 models:
#ERROR METRICS
m312css <- accuracy(m312$css)[1:7]
m012css <- accuracy(m012$css)[1:7]
m013css <- accuracy(m013$css)[1:7]
m112css <- accuracy(m112$css)[1:7]
m113css <- accuracy(m113$css)[1:7]
m0114css <- accuracy(m0114$css)[1:7]
df.Smodels <- data.frame(
  rbind(m312css, m012css, m013css, m112css, m113css, m0114css)
)
colnames(df.Smodels) <- c("ME", "RMSE", "MAE", "MPE", "MAPE",
                          "MASE", "ACF1")
rownames(df.Smodels) <- c("SARIMA(3,1,2)x(0,1,1)_12", "SARIMA(0,1,2)x(0,1,1)_12",
"SARIMA(0,1,3)x(0,1,1)_12",
                          "SARIMA(1,1,2)x(0,1,1)_12", "SARIMA(1,1,3)x(0,1,1)_12",
"SARIMA(0,1,14)x(0,1,1)_12"
)
round(df.Smodels,  digits = 3)


#Residuals of best models from BIC still have
#significant autocorrelation
residual.analysis(model = m012$css)
#Best models from AIC captured the information well
#and nothing valuable is in their residuals

m1114 = fit_SARIMA(sqrt_data,orders=c(1,1,14)) #About half of the vars are significant
residual.analysis(m1114$ml) #Capture info pretty well

#Try overfitted model
m0115 = fit_SARIMA(sqrt_data,orders=c(0,1,15)) #Most vars are significant in CSS
residual.analysis(m0115$ml) #Capture info very well
#ERROR METRICS
m312css <- accuracy(m312$css)[1:7]
m012css <- accuracy(m012$css)[1:7]
m013css <- accuracy(m013$css)[1:7]
m112css <- accuracy(m112$css)[1:7]
m113css <- accuracy(m113$css)[1:7]
m0114css <- accuracy(m0114$css)[1:7]
m0115css <- accuracy(m0115$css)[1:7]
m1114css <- accuracy(m1114$css)[1:7]
```

```
df.Smodels <- data.frame(
  rbind(m312css, m012css, m013css, m112css, m113css, m0114css, m0115css,m1114css)
)
colnames(df.Smodels) <- c("ME", "RMSE", "MAE", "MPE", "MAPE",
                          "MASE", "ACF1")
rownames(df.Smodels) <- c("SARIMA(3,1,2)x(0,1,1)_12", "SARIMA(0,1,2)x(0,1,1)_12",
"SARIMA(0,1,3)x(0,1,1)_12",
                          "SARIMA(1,1,2)x(0,1,1)_12", "SARIMA(1,1,3)x(0,1,1)_12",
"SARIMA(0,1,14)x(0,1,1)_12",
                          "SARIMA(0,1,15)x(0,1,1)_12","SARIMA(1,1,14)x(0,1,1)_12")
round(df.Smodels,  digits = 3)


sc.AIC = AIC(m312$ml,m012$ml,m013$ml,m112$ml,m113$ml,m0114$ml,m0115$ml,m1114$ml)
sc.BIC = BIC(m312$ml,m012$ml,m013$ml,m112$ml,m113$ml,m0114$ml,m0115$ml,m1114$ml)
sort.score(sc.AIC, score = "aic")
sort.score(sc.BIC, score = "bic")
#Big models capture information well but have the worst aic and bic scores

#AIC picks m316 over 315

#FORECASTING


FC_CSS = Arima(subset,order=c(0,1,15),seasonal=list(order=c(0,1,1), period=12),
               lambda = 0.5, method = "CSS")
forecastCSS = forecast(FC_CSS,lambda = 0.5,  h = 10)
forecastCSS
plot(forecastCSS)

FC_ML = Arima(subset,order=c(0,1,15),seasonal=list(order=c(0,1,1), period=12),
                lambda = 0.5, method = "ML")
forecastML = forecast(FC_ML,lambda = 0.5,  h = 10)
forecastML
plot(forecastML)
```