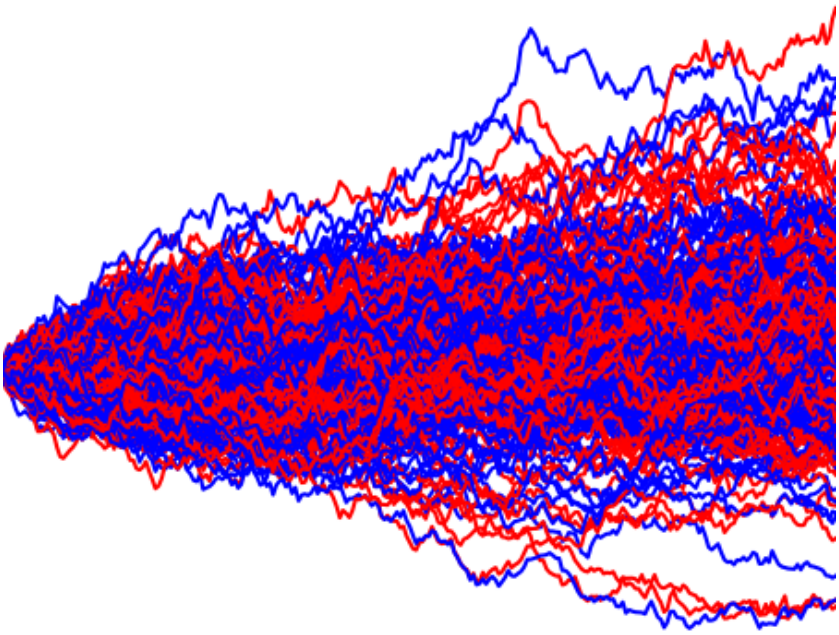
YOUTH EMPLOYMENT PREDICTION
MaBosso



Authors:
Ditiro Letsoalo
Omolemo Mompati

Student Numbers:
LTSDIT002
MMPOMO001

August 2024

Introduction

The South African Youth Employment Prediction Challenge is a data-driven competition aimed at addressing one of the most pressing issues in South Africa: youth unemployment. By analyzing comprehensive labour force data from January 2008 to June 2024, participants are tasked with developing predictive models to forecast the number of employed males and females, aged 18 to 34, in each province for the third quarter of 2024. This challenge not only seeks to enhance predictive accuracy but also aims to extract actionable insights that can inform policies to boost youth employment. Through this initiative, we hope to contribute valuable knowledge to policymakers, researchers, and businesses, ultimately fostering a more prosperous future for South Africa's youth.

Dataset Description

The dataset provided for this challenge spans from January 2008 to June 2024 and includes detailed quarterly data on employment trends across different provinces in South Africa. The data captures various attributes related to employment, such as the number of employed males and females aged 18 to 34, across all provinces. This rich dataset offers a comprehensive view of the labour market dynamics over a significant period, enabling robust analysis and predictive modeling.

Data Exploration

Initial exploratory data analysis (EDA) reveals several key trends and patterns in the dataset:

- **Temporal Trends**: Employment numbers exhibit seasonal variations and long-term trends, reflecting economic cycles and policy impacts.
- **Provincial Differences**: There are notable differences in employment trends across provinces, highlighting regional disparities in economic opportunities and challenges.
- **Gender Disparities**: The data shows variations in employment rates between males and females, which may be influenced by socio-economic factors and policy interventions.

Data Cleaning

To ensure the quality and reliability of the dataset, several data cleaning steps were undertaken:

- **Handling Missing Values**: Missing values were addressed using appropriate imputation techniques or by excluding incomplete records, depending on the extent and nature of the missing data.
- **Correcting Inconsistencies**: Inconsistencies in the data, such as duplicate records or incorrect entries, were identified and corrected to maintain data integrity.

Data Splitting

The dataset was split into training and test sets to ensure robust model development and evaluation. The training set, comprising the majority of the quarterly data from January 2008 to June 2024, was used to train the predictive models. The test set, containing data for the third quarter of 2024, was used to evaluate the final model's performance, providing an unbiased assessment of its predictive accuracy. This approach ensures that the model is trained on a comprehensive dataset while being tested on the next quarter's data to gauge its effectiveness.

Data Visualisation

Table 1: Summary Statistics for Important Numerical Variables

Variable	Count	Mean	Std Dev	Min	25%	50%	75%	Max
AGE	1,305,343	25.51	4.90	18.00	21.00	25.00	30.00	34.00
TIME UNEMPLOYED	712,329	1.40	2.35	0.00	0.00	0.00	2.00	99.00
Weight	1,305,343	826.87	530.35	44.64	490.48	715.69	1,012.45	17,356.19

The summary statistics in the above table provides an overview of three key numerical variables: AGE, TIME UNEMPLOYED, and Weight. The average age of individuals is 25.51 years, with most being between 21 and 30 years old. The average time unemployed is 1.40 years, with a median of 0 years, indicating that more than half of the individuals have not experienced unemployment. The average weight is 826.87 units, with a wide range of values, suggesting significant variability. Notably, there are some extreme values, particularly in the TIME UNEMPLOYED and Weight variables, which may indicate outliers or data entry errors.

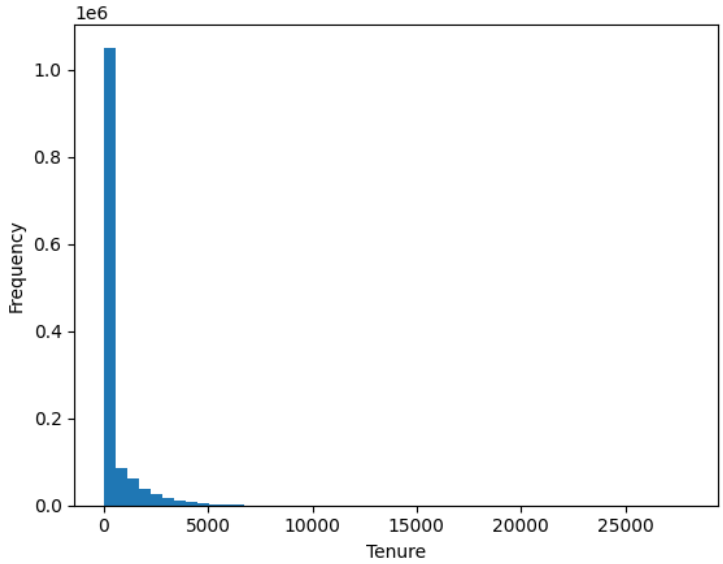


Figure 1: The Histogram showing the Distribution of Tenure.

The histogram of tenure values reveals a right-skewed distribution, with the majority of values clustered around zero. This indicates that most individuals in the dataset have very short employment durations. The presence of a few high tenure values suggests that only a small number of individuals maintain long-term employment. This pattern is significant for understanding youth employment trends in South Africa, as it highlights the challenges young people face in securing stable, long-term jobs. The skewed distribution implies that while a few individuals may achieve extended employment and potentially higher earnings, the majority experience short-term employment, which could limit their earning potential and career growth. These insights underscore the need for targeted policies that promote job stability and long-term employment opportunities for the youth, such as skills development programs, job matching services, and incentives for employers to retain young employees.

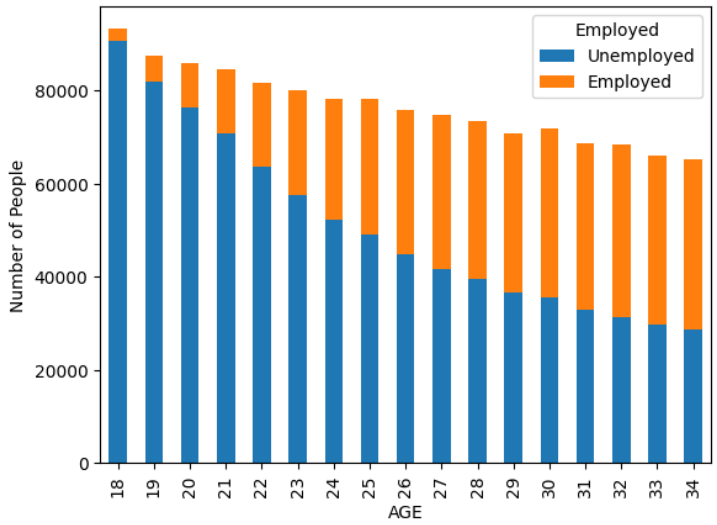


Figure 2: The Number of Employed and Unemployed people by Age.

The above chart shows that older individuals are generally more likely to be employed compared to younger individuals. This trend is particularly significant for the youth employment data you're analyzing. The younger age groups (18-24 and 25-34) have higher unemployment rates compared to older age groups. This disparity highlights the challenges faced by young people in securing stable employment in South Africa. in this case visualization underscores the importance of developing targeted policies to address youth unemployment. Initiatives could include enhancing vocational training, creating more entry-level job opportunities, and providing incentives for businesses to hire and retain young employees. By focusing on these areas, policymakers can help bridge the employment gap and improve job stability for the youth.

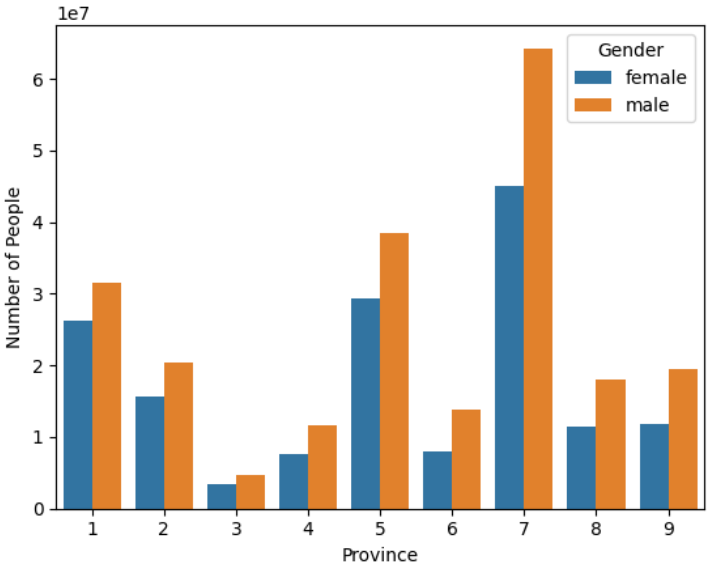


Figure 3: Number of Employed People per Gender and Province.

The above plot shows that in Province 1, the number of employed males might be significantly higher than the number of employed females, indicating a gender disparity in employment. Conversely, in Province 2, the number of employed females could be higher than that of males, suggesting a different trend. This

pattern continues across all provinces, with some showing a balanced employment distribution between genders, while others exhibit noticeable differences. By examining these numbers, researchers can identify specific provinces where gender-based employment inequalities are more pronounced. This information is crucial for policymakers and organizations aiming to address these disparities and promote gender equality in the workforce. Understanding these trends can help in developing targeted interventions to support underrepresented groups and ensure equitable employment opportunities across all regions.

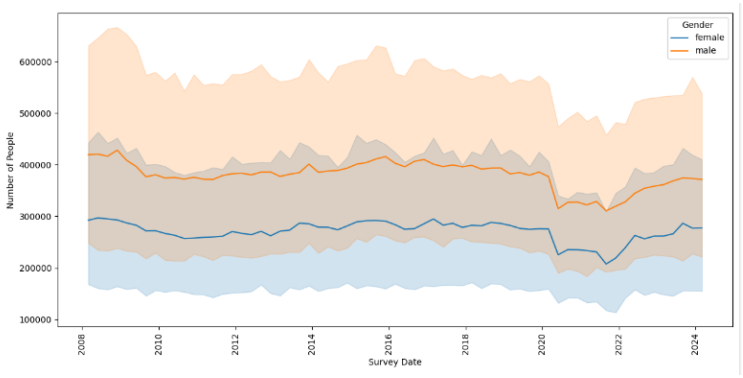


Figure 4: The Times series showing number of Employed people per Province.

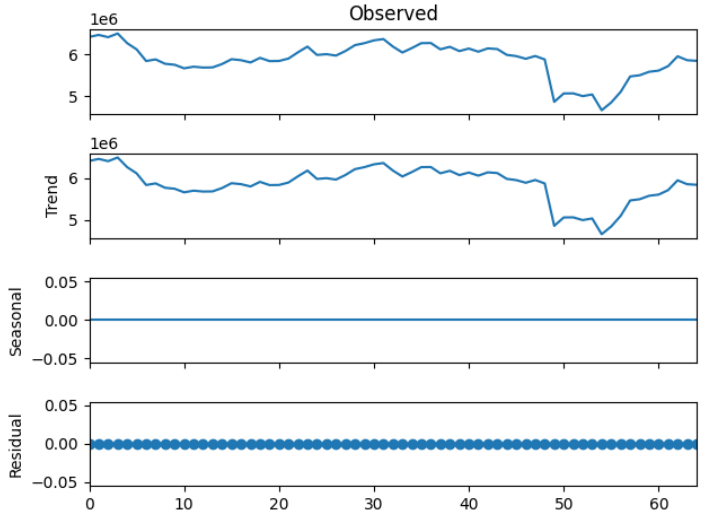


Figure 5: Seasonal Decomposition of Employment Data: Additive Model Analysis, Showing Trend, Seasonal, and Residual Components.

The time series decomposition plot provided above breaks down youth employment data into observed, trend, seasonal, and residual components. The observed component shows the actual data over time, while the trend component highlights the long-term progression. The seasonal component, appearing as a flat line, indicates no significant seasonal effects. The residuals display random variations around a zero line, representing anomalies not explained by trend or seasonality. This decomposition is crucial for understanding the underlying patterns in youth employment, identifying trends, and making accurate predictions. The presence of a trend suggests non-stationarity, as the statistical properties of the series change over time.

0.1 Data Differencing

In time series analysis, **data differencing** is a crucial technique used to stabilize the mean of a time series by removing changes in the level of the series, which can be caused by trends or seasonality. After examining the data, we observed a noticeable trend, indicating that the mean value of the series is not constant over time. This violates one of the key assumptions of many time series forecasting models, which require stationarity. To address this issue, we applied first-order differencing, which involves subtracting the previous observation from the current observation. Mathematically, this can be represented as:

$$Y_t' = Y_t - Y_{t-1}$$

where Y_t' is the differenced value at time t, Y_t is the original value at time t, and Y_{t-1} is the value at the previous time point.

1 Methodology

In this study, we utilize a variety of time series forecasting models, including Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), and Seasonal ARIMA (SARIMA). These models are selected based on their suitability for analyzing the employment data and their ability to capture underlying trends and seasonal patterns.

The primary assumptions underlying these models include linearity, stationarity, and the absence of multicollinearity. It is essential that the time series data adheres to these assumptions to ensure the validity of the forecasts produced by these models.

Although models such as ARCH (Autoregressive Conditional Heteroskedasticity) and GARCH (Generalized Autoregressive Conditional Heteroskedasticity) are effective for time series data characterized by volatility clustering, they are not appropriate for our analysis. The employment data exhibits little to no volatility, making the use of ARCH and GARCH models unnecessary. Therefore, we focus on AR, MA, ARMA, ARIMA, and SARIMA to accurately capture the trends and patterns in the data.

Autoregressive (AR)

Autoregressive is a statistical technique used in time-series analysis that assumes that the current value of a time series is a function of its past values. Autoregressive models use similar mathematical techniques to determine the probabilistic correlation between elements in a sequence and it is used for analyzing and forecasting time series data.

An AR model of order p, denoted as AR(p), can be mathematically expressed as:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t$$

where X_t is the current value of the time series, c is some constant and $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the model that represent the influence of previous values. ϵ_t is a white noise error term that follows a normal distribution with $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ where the errors are independent and identical.

Moving Average (MA)

A Moving Average (MA) model is a statistical method used in time series analysis to smooth out short-term fluctuations and highlight longer-term trends or cycles. An MA model of order q, denoted as MA(q), expresses the current value of a time series as a linear combination of past error terms (shocks). Mathematically, it can be represented as:

$$X_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t$$

where X_t is the current value of the time series, $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the model that represent the influence of previous error terms and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is a white noise error term at time t. The errors are independent and identical.

Autoregressive Moving Average (ARMA)

An ARMA (Autoregressive Moving Average) model is a popular statistical model used for analyzing and forecasting time series data. It combines the features of both autoregressive and moving average models. An ARMA model of order (p, q), denoted as ARMA(p, q), is defined as:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t$$

where:

- X_t is the current value of the time series.
- c is a constant (intercept).
- $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the autoregressive component.
- $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the moving average component.
- $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is a white noise error term at time t. The errors are identical and independent.

Autoregressive Moving Average (ARMA)

An ARMA (Autoregressive Moving Average) model is a popular statistical model used for analyzing and forecasting time series data. It combines the features of both autoregressive and moving average models. An ARMA model of order (p, q), denoted as ARMA(p, q), is defined as:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t$$

where:

- X_t is the current value of the time series.
- c is a constant (intercept).
- $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the autoregressive component.
- $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the moving average component.
- $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is a white noise error term at time t. The errors are identical and independent.

Autoregressive Integrated Moving Average (ARIMA)

The Autoregressive Integrated Moving Average model is a widely used statistical method for forecasting time series data, particularly when the data exhibits non-stationarity. An ARIMA model is denoted as ARIMA(p, d, q).

The Differences of order d are defined as $\nabla^d = (1 - B)^d$. The model can be expressed as:

$$\phi(B)(1 - B)^d X_t = \theta(B)\epsilon_t$$

where

$$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p, \theta(B) = (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q)$$

such that for $B^h X_t = X_{t-h}$ for $h \in \mathbb{Z}$. X_t is the current value of the time series, $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the autoregressive component, $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the moving average component and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is a white noise error term at time t.

Seasonal Autoregressive Integrated Moving Average (SARIMA)

The Seasonal Autoregressive Integrated Moving Average model is an extension of the ARIMA model that incorporates seasonality in time series data.

A SARIMA model is denoted as $ARIMA(p, d, q)(P, D, Q)_m$, The model can be expressed as:

$$\Phi(B^m)\phi(B)(1 - B^m)^D(1 - B)^d X_t = \Theta(B^m)\theta(B)\epsilon_t$$

where

$$\begin{aligned}
\phi(B) &= (1 + \phi_1 B + \phi_2 B^2 + \ldots + \phi_p B^p, \\
\theta(B) &= (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q), \\
\Phi(B) &= (1 - \Phi_1 B^m - \Phi_2 B^{2m} - \ldots - \Phi_P B^{Pm}, \\
\theta(B) &= (1 + \Theta_1 B + \Theta_2 B^2 + \ldots + \Theta_Q B^{Qm})
\end{aligned}$$

such that ϕ_i and Φ_k are the parameters of the autoregressive components, θ_j and Θ_m are the parameters of the moving average components and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is a white noise error term at time t.

2 Results

To determine the appropriate models for our time series analysis, we examined the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the employment data. These tools are instrumental in identifying the underlying structure of the time series and guiding the selection of suitable ARIMA and SARIMA models.

The ACF measures the correlation between observations at different lags, providing insights into the extent of correlation between a time series and its past values. A rapid decay in the ACF plot suggests that an MA model may be appropriate, while a slow decay can indicate the presence of an AR model. The PACF, on the other hand, measures the correlation between an observation and its lagged values after removing the effects of intervening observations. A significant cut-off in the PACF plot indicates the order of the AR component in the model.

Given the diversity in the dataset, I used different models for various combinations of gender and province. This approach recognizes that employment trends may differ significantly across these demographics. By analyzing the ACF and PACF plots for each combination, we could tailor our model selection to fit the specific characteristics of the data, ensuring a more accurate representation of the underlying patterns.

While selecting models, we were mindful of the balance between underfitting and overfitting. Even though we observed many significant spikes in the ACF plots a lot, it is crucial to avoid fitting overly complex models that could lead to overfitting. Overfitting occurs when a model captures noise rather than the underlying signal, resulting in poor predictive performance on unseen data. Conversely, underfitting can happen when a model is too simplistic to capture essential features of the data.

To mitigate these risks, we employed model validation techniques, such as cross-validation (RMSE), to assess model performance. This careful approach allowed us to select models that not only fit the training data well but also generalized effectively to new observations, ultimately enhancing the reliability of our forecasts for employment trends across different gender and province combinations.

To start with, we differenced the time series plots as there were not stationary due to the trend. The following tests were then performed to check stationarity:

ADF Test

The **Augmented Dickey-Fuller (ADF) test** is used to determine whether a unit root is present in a univariate time series. The null and alternative hypothesis are given as follows:

$$H0: \text{ The time series has a unit root.}$$
$$H1: \text{ The time series does not have a unit root.}$$

KPSS Test

The **Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test**, in contrast, tests the null hypothesis that the time series is stationary around a deterministic trend. The hypothesis is done as follows:

$$H0: \text{ The time series is stationary.}$$
$$H1: \text{ The time series is non-stationary.}$$

The results of the gender on each province is given as following table:

Table 2: ADF and KPSS Test Results by Province and Gender

Province	**Gender**	**ADF p-value**	**KPSS p-value**
1	male	4.0617e-04	0.1000
1	female	1.2713e-11	0.1000
2	male	3.8277e-18	0.1000
2	female	1.3370e-02	0.1000
3	male	5.2786e-11	0.1000
3	female	1.1844e-14	0.1000
4	male	9.8689e-16	0.1000
4	female	1.4947e-06	0.1000
5	male	1.2013e-13	0.1000
5	female	8.3316e-11	0.1000
6	male	1.7464e-17	0.1000
6	female	2.6985e-11	0.1000
7	male	2.5412e-09	0.1000
7	female	6.7005e-15	0.1000
8	male	1.3228e-14	0.1000
8	female	4.5848e-09	0.1000
9	male	8.6918e-18	0.1000
9	female	2.9431e-03	0.0598

For the ADF test, a small p-value (p-value < 0.05) indicates that we reject the null hypothesis (H0) while for the KPSS test, a small p-value (p-value > 0.05) indicates that we reject the null hypothesis (H0). Based on the above table, the results indicate that the time series across all provinces and genders are stationary based on both the ADF and KPSS tests since in the ADF tests all the p-values re small and KPSS's p-values are large.

The Models

Following the presentation of the ADF and KPSS test results by province and gender in above table , we proceeded to fit models for the time series data based on the insights gained from the autocorrelation function (ACF) and partial autocorrelation function (PACF). We analyzed the ACF and PACF plots for each gender within each province to identify the appropriate model structures. These plots help determine the order of the autoregressive (AR) and moving average (MA) components in the time series models. By examining the significant lags in the ACF and PACF, we could ascertain the optimal parameters needed for model fitting. In some other models SARIMA models were used as there is little seasonality. The ACF and PACF plots are in the code which can be accessed.

Data Summary

The following table presents the predicted versus actual employment values along with the corresponding keys:

Table 3: Predicted vs Actual Employment Values

yhat	**actual**	**key**
478239	478188.275720	(1, female)
566974	566490.924327	(1, male)
235703	240100.969020	(2, female)
283370	282952.210806	(2, male)
50766	50690.941062	(3, female)
69577	69632.786320	(3, male)
84830	84734.857333	(4, female)
151428	151784.723075	(4, male)
459649	460789.634325	(5, female)
643394	647106.667222	(5, male)
110454	95701.414543	(6, female)
193772	192301.470470	(6, male)
687632	681222.306926	(7, female)
893500	853611.452875	(7, male)
195050	199023.189482	(8, female)
252854	252768.903376	(8, male)
202988	207631.298348	(9, female)
331500	326681.836255	(9, male)

The table presents a comparison of predicted employment values (yhat) against actual figures for different provinces and genders in South Africa, revealing a generally strong performance of the predictive model. Most predicted values are close to actual figures, suggesting reliability; however, notable discrepancies exist, particularly with (6, female), where the prediction significantly underestimates the actual value. While many entries reflect a conservative approach, indicating the model's cautiousness, these deviations highlight areas for potential improvement. Further analysis is needed to understand the underlying factors contributing to these discrepancies, particularly for demographic groups that may be underrepresented in the model. Overall, the model shows promise but requires continuous refinement to enhance its accuracy in predicting employment trends.