MA641 Final Project

Akanksha Wagh

Spring 2025

Project 2: Non-Seasonal Dataset

Title: Forecasting Metro-Level Hiring Trends

1. INTRODUCTION:

Metropolitan labor market dynamics are important markers of worker demand. We model and estimate employment changes over time in this study by looking at monthly job posting data provided by Indeed at the metro level in the United States, particularly in New York City. This forecast seeks to uncover trends in job demand among New York Metro cities. By utilizing time series analytic methods including ARIMA model selection, ACF/PACF evaluation, and stationarity testing, we can pinpoint important features in the data.

2. DATA DESCRIPTION:

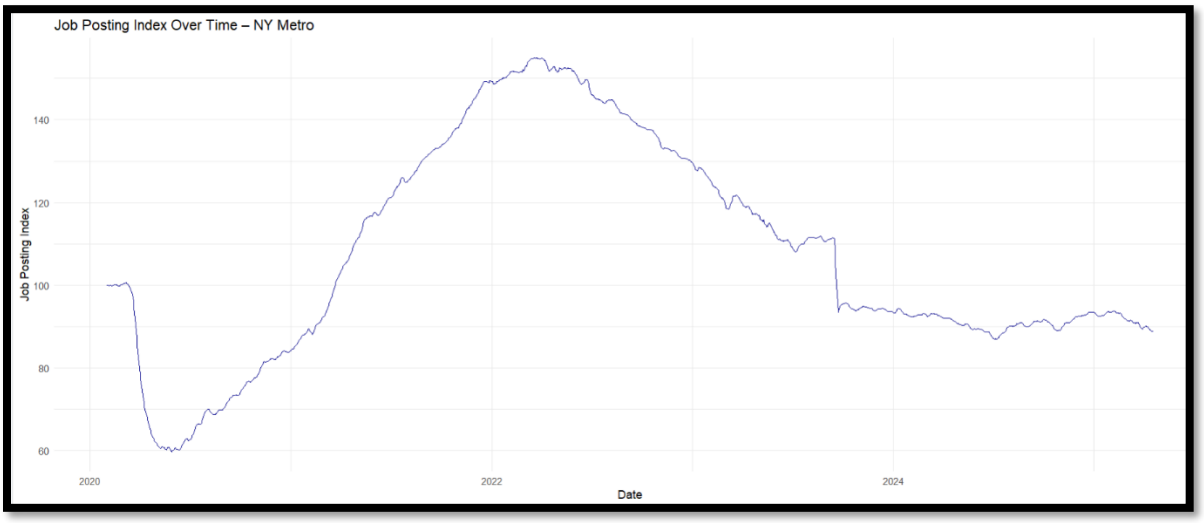
Time Range: February 1, 2020 to February 28, 2025 (Daily)

Columns:

- date: Daily timestamp of the job posting index
- metro: Metropolitan Statistical Area (MSA), here "Abilene, TX"
- cbsa_code: Core-Based Statistical Area code identifying the region (e.g., 10180 for Abilene, TX)
- indeed_job_postings_index: Indexed job posting activity with base value of 100 on February 1, 2020

3. ANALYSIS OF INITIAL OTP DATASET:

Daily Job Posting Performance Over Time – NY Metro:



Sudden dip due to the COVID-19 pandemic can be observed from the above graph. Significant rise throughout 2021–2022. Steady decline in job postings after 2022. Stabilization but at lower index levels since last year.

4. STATIONARITY TEST:

Stationarity is an important factor when it comes to time series analysis because the stability in data makes it easier for the models to learn and extrapolate patterns. Models like ARIMA and SARIMA consider the data to be stationary, and the estimates can be unreliable or biased. I used Augmented Dickey-Fuller (ADF Test) to check the stationarity of the series.

Ho: Series is non-stationary.

If the p-value < 0.05 then, we reject Ho which means our data is stationary and we can move ahead. But if the p-value > 0.05, we fail to reject Ho which means the data is non-stationary.

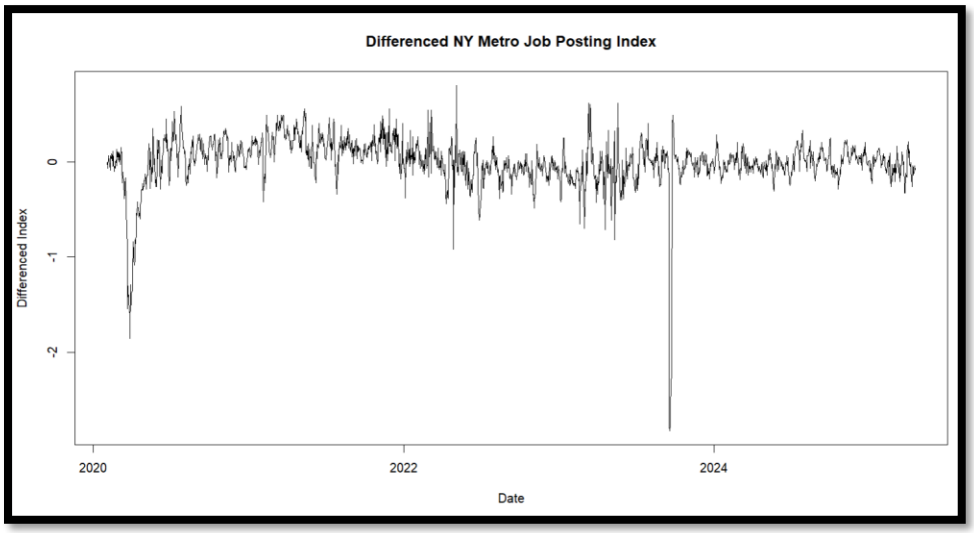
```
        Augmented Dickey-Fuller Test

data:  ny_vector
Dickey-Fuller = -0.87429, Lag order = 12, p-value = 0.955
alternative hypothesis: stationary
```

As we can see the p-value is 0.955. This means we fail to reject Ho, and our data is non-stationary.

5. DIFFERENCING:

Differencing is used for non-stationary data because such data often shows trends or fluctuations in mean over time. Differencing helps to remove these trends, effectively flattening any upward or downward movement in the series and making the mean constant over time.



This transformation is important for maintaining the stationarity condition required by ARIMA, which assumes the data has stable statistical properties. Differencing prepares the time series for more accurate and reliable modeling.

After differencing, we do the ADF Test again to check the stationarity.

Ho: Series is non-stationary.

If the p-value < 0.05 then, we reject Ho which means our data is stationary and we can move ahead. But if the p-value >0.05, we fail to reject Ho which means the data is non-stationary.

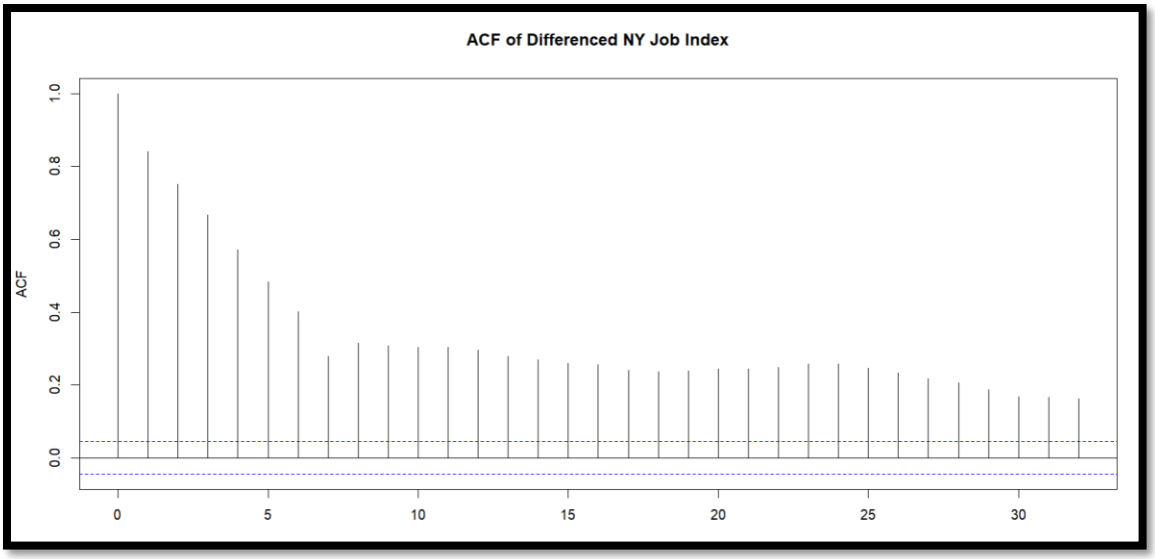
```
        Augmented Dickey-Fuller Test

data:  as.numeric(ny_diff1)
Dickey-Fuller = -7.7169, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary
```

As we can see the p-value is 0.01. This means we reject Ho and our data is stationary.

6. ACF & PACF PLOTS:

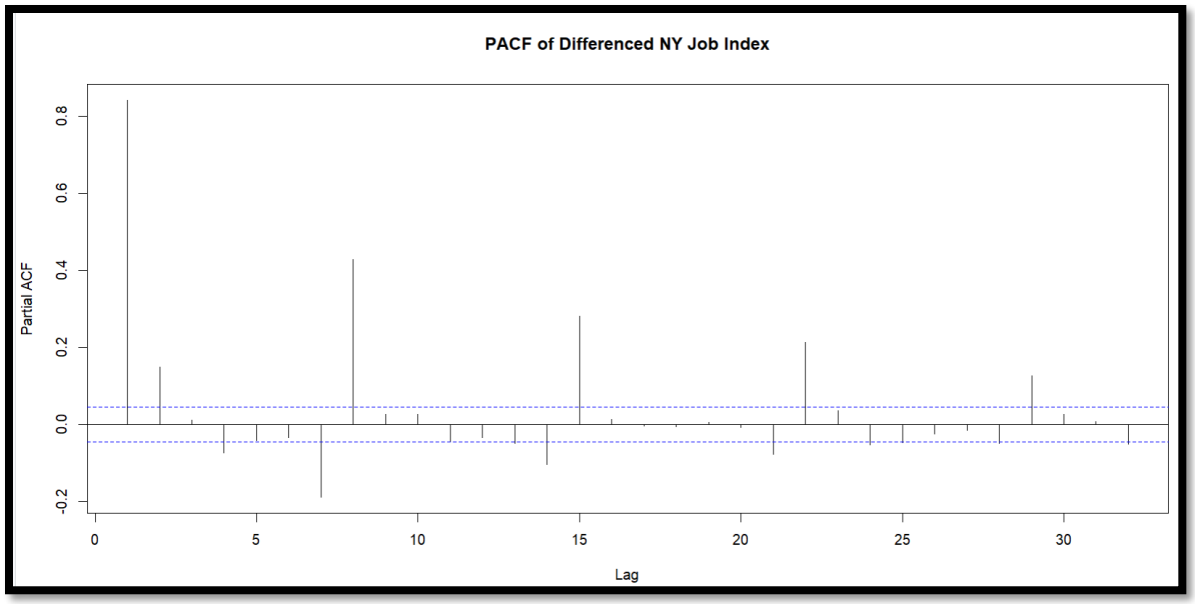
Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots are crucial diagnostic tools for time series forecasting because these plots help us understand the structure and pattern of our data and directs us to choose the best model – especially in the case of ARIMA & SARIMA.



Let's suppose we are selecting parameter for the model ARIMA(p, d, q); **ACF** helps you choose **q (MA order)**, **PACF** helps you choose **p (AR order)** AND **d** is decided by differencing

to make the series stationary.

If we observe our ACF Plot, there is gradual decay which indicates MA(p) component. In PACF Plot, there is sharp cut off at lag 1, which indicate AR(1).



From the plots, we will consider the following ARIMA model candidates:

ARIMA (1,1,1), ARIMA (0,1,1) or ARIMA (2,1,1)

7. MODEL SELECTION & COMPARISON:

Since we have 3 ARIMA models, we will compare them using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). We use AIC and BIC to select the best models for prediction, simplicity and interpretability. The model which has the lowest AIC and BIC is the best for the dataset we are working on.

```
> print(model_comparison)
         Model        AIC        BIC
1 ARIMA(0,1,1)  -661.4825  -650.3802
2 ARIMA(1,1,1) -1861.8983 -1845.2447
3 ARIMA(2,1,1) -1870.2418 -1848.0370
>
```

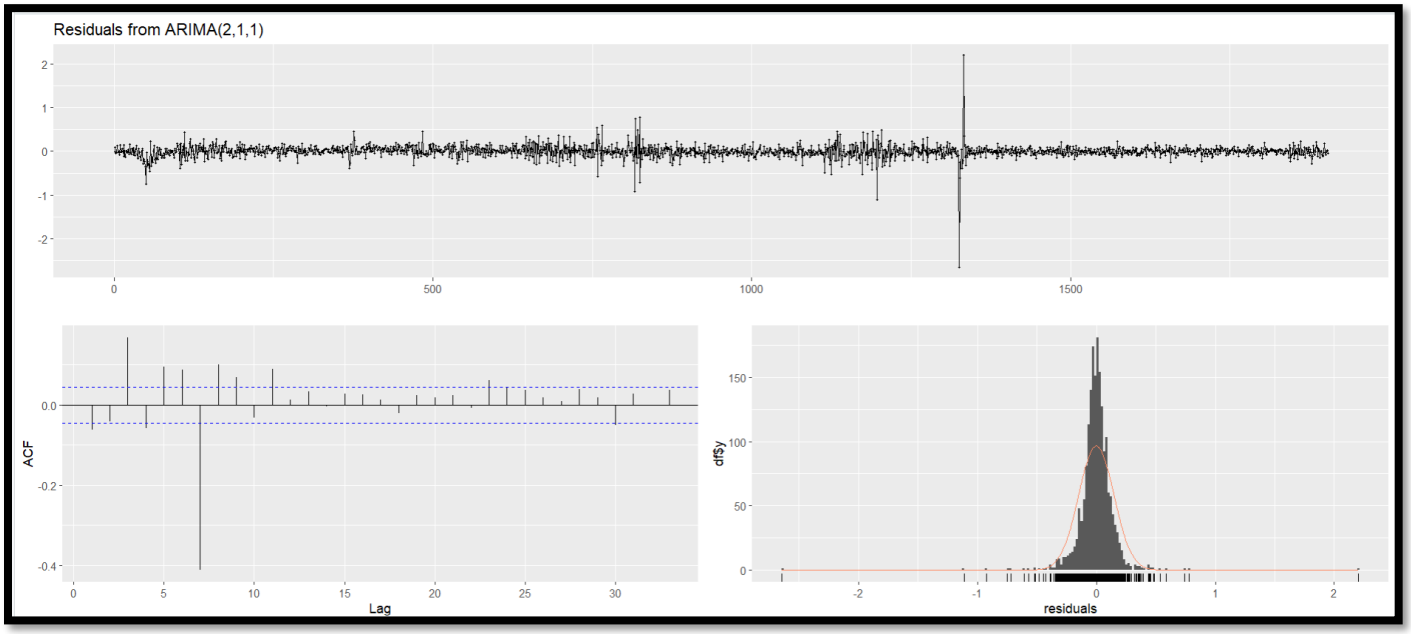
ARIMA (2,1,1) has the lowest values of AIC & BIC i.e., -1870.2418 & -1848.0370 respectively. It suggests balance in model complexity, and it is the best fit.

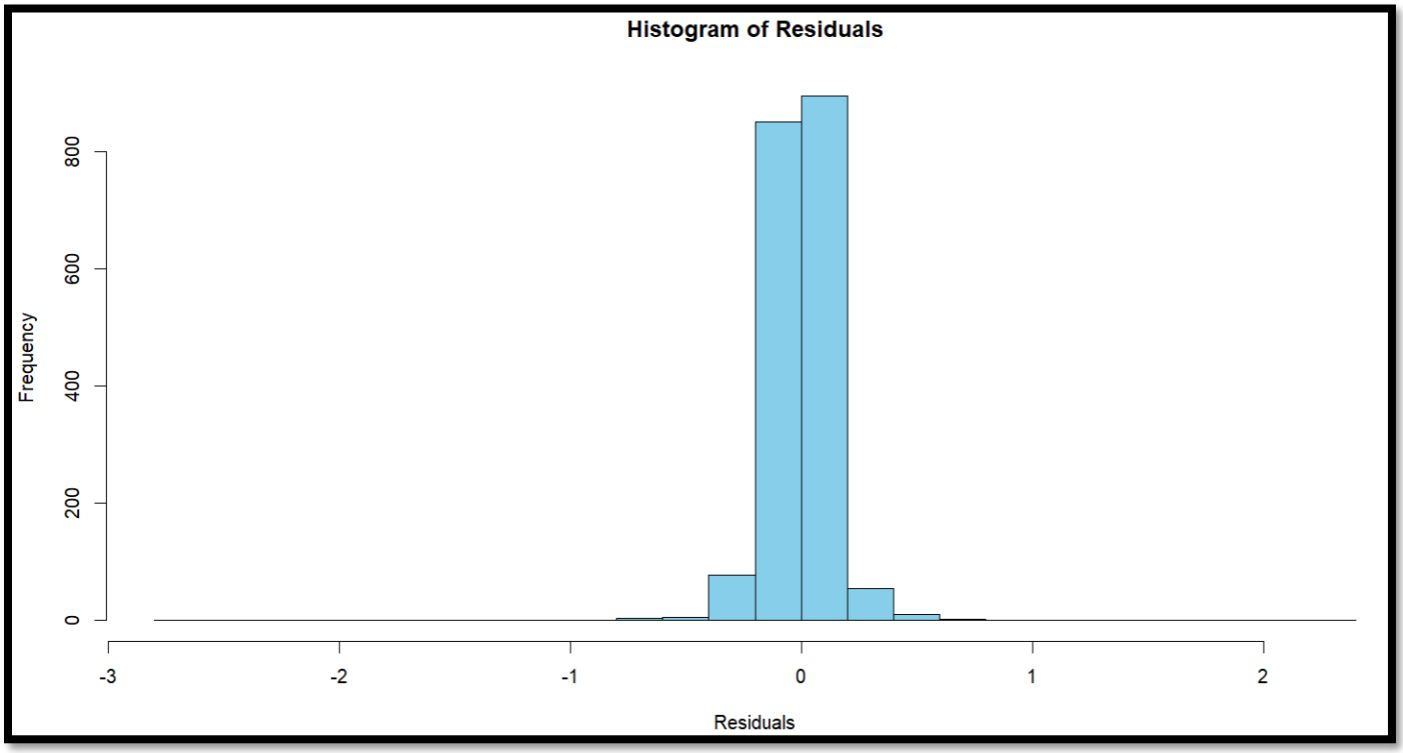
8. RESIDUAL DIAGNOSTICS & MODEL FITTING:

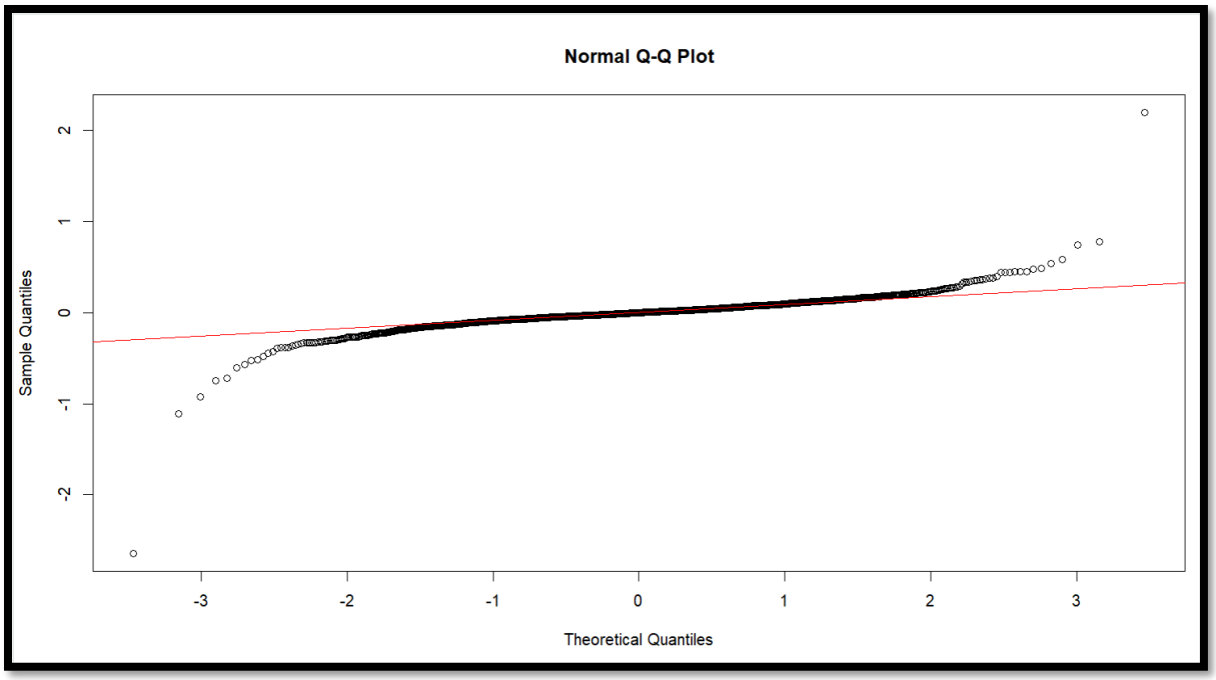
Even if the model is selected, we will perform Residual Diagnostic. It is important because it checks whether the selected model is model is appropriate, reliable, and avoids overfitting or underfitting. What are residuals? Residual is the error between actual and predicted values which should be random, uncorrelated and normally distributed. If any one condition is not satisfied that means, there were patterns which were not captured. We will check the

autocorrelation by plotting ACF of residuals, normality by QQ-Plot and Histogram, white noise by Ljung-Box Test.

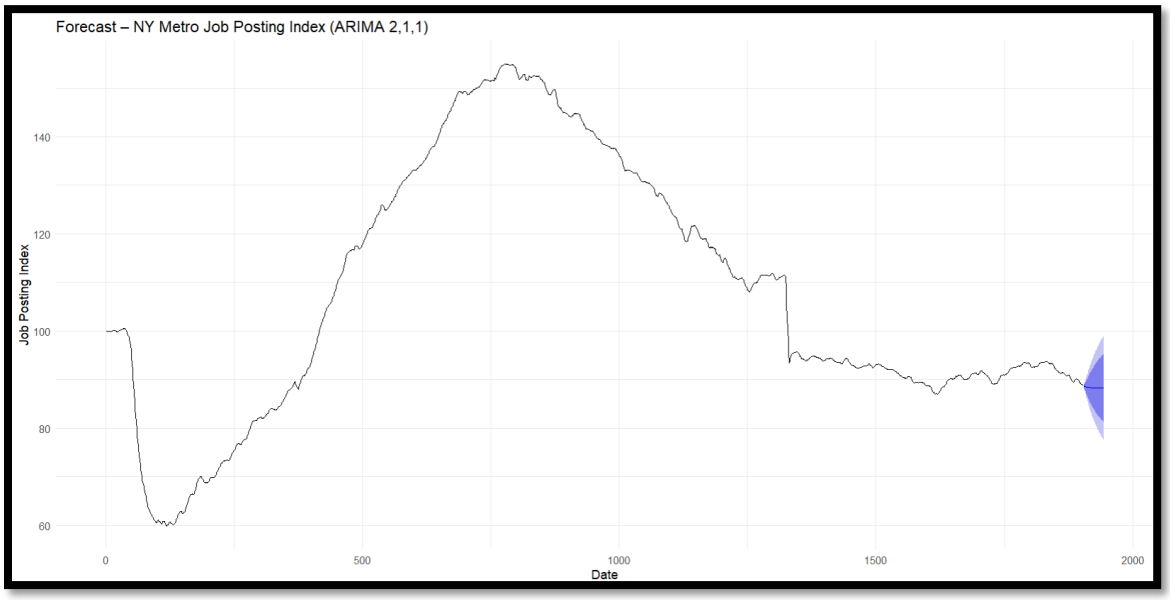
In our case, the residuals plot fluctuates around zero which is good and there's no visible trend. The ACF of Residuals spikes within the confidence line. The histogram is bell-shaped, indicating the normal distribution. QQ-plot has minor deviation from the line but all together it is a sign of normality. Moving ahead with the Ljung-Box Test, the p-value is 0.347 which means the p-value is much greater than 0.05, indicating significant autocorrelation in residuals and gives white-noise residuals.







9. FORECASTING FOR THE NEXT 12 MONTHS:

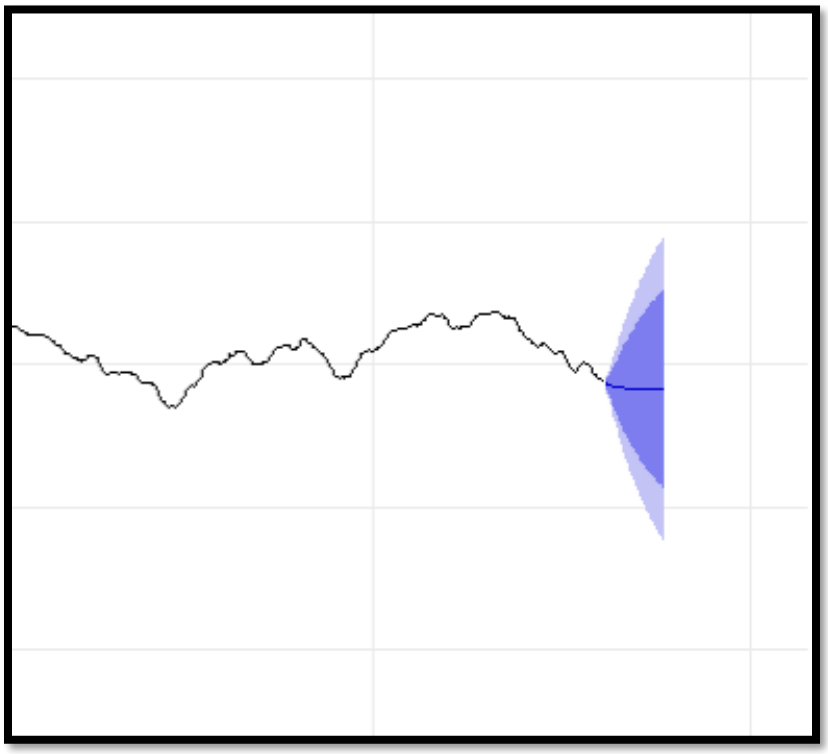


The blue shaded part of the graph is the Job Posting forecast for the next 40 days. The dark blue part of the forecast is 80% confidence interval whereas the light blue part of the forecast is 95% confidence interval.

10. CONCLUSION:

The dataset shows no trends. It had a sudden dip during COVID-19, and the job market rose immediately within months. At the end of the forecast, you can see there is a significant rise

or dip, but the data has a good range to maintain stability in the markets for at least the next 40 days.



APPENDIX [Code]

```
#Non-Seasonal Dataset of Indeed Job Postings all over US.
#Forecasting the On-Time Performance for net 12 months
library(tidyverse)
library(lubridate)
library(ggplot2)
library(zoo)
library(urca)
library(tseries)
library(forecast)

setwd("C:/Users/akank/Downloads")
data <- read.csv("metro_job_postings_us.csv", stringsAsFactors = FALSE)

#Cleaning the data and converting the data type
names(data) <- toupper(trimws(names(data)))
data$DATE <- as.Date(data$DATE)
head(data, 10)
#List of unique column values
colnames(data)
unique(data$METRO)

# GG Plot only for job postings for New York
ggplot(data %>% filter(METRO == "New York-Newark-Jersey City, NY-NJ-PA"), aes(x = DATE, y
= INDEED_JOB_POSTINGS_INDEX)) + geom_line(color = "darkblue") +
  labs(title = "Job Posting Index Over Time - NY Metro", x = "Date", y = "Job Posting
Index") + theme_minimal()

# Further project is working only for New York metro
ny_data <- data %>% filter(METRO == "New York-Newark-Jersey City, NY-NJ-PA")
ny_data <- ny_data %>% arrange(DATE)

#Time series object for Job Postings in NY
ny_ts <- zoo(ny_data$INDEED_JOB_POSTINGS_INDEX, order.by = ny_data$DATE)
plot(ny_ts, main = "NY Metro Job Posting Index (Zoo Time Series)", ylab = "Index", xlab =
"Date")

# Convert to numeric vector for tests
ny_vector <- as.numeric(ny_ts)
# Check Stationarity

#ADF Test (Augmented Dickey-Fuller)
#H0: Series is non-stationary.(If p-value<0.05 -> Reject H0)
adf.test(ny_vector)

# In ADF test our OTP Time series is non-stationary.
# First-order differencing
ny_diff1 <- diff(ny_ts)
plot(ny_diff1, main = "Differenced NY Metro Job Posting Index", ylab = "Differenced
Index", xlab = "Date")

# Check stationarity tests
#H0: Series is non-stationary.(If p-value<0.05 -> Reject H0)
adf.test(as.numeric(ny_diff1))

# In ADF test our data is now stationary after differencing.
# There are few spikes but not enough to invalidate modeling.

# ACF and PACF plots of differenced job posting series
# AR (p): At PACF - cut-off at lag p, gradual decay in ACF.
# MA (q): At ACF - cut-off at lag q, gradual decay in PACF.
acf(ny_diff1, main = "ACF of Differenced NY Job Index")
pacf(ny_diff1, main = "PACF of Differenced NY Job Index")

#MA(1) AND AR(2)
```

```
#ARIMA Models: ARIMA(0,1,1) , ARIMA(1,1,1), ARIMA(2,1,1)
model_011 <- Arima(ny_vector, order = c(0,1,1))
model_111 <- Arima(ny_vector, order = c(1,1,1))
model_211 <- Arima(ny_vector, order = c(2,1,1))

# Compare models using AIC and BIC
model_comparison <- data.frame(
  Model = c("ARIMA(0,1,1)", "ARIMA(1,1,1)", "ARIMA(2,1,1)"),
  AIC = c(AIC(model_011), AIC(model_111), AIC(model_211)),
  BIC = c(BIC(model_011), BIC(model_111), BIC(model_211))
)
print(model_comparison)

#Model ARIMA(2,1,1)

# Residual diagnostics
checkresiduals(model_211)
residuals_211 <- residuals(model_211)
ny_ts <- ts(ny_vector, start = c(2020, 1), frequency = 12)

# Histogram for Normality
hist(residuals_211, main = "Histogram of Residuals", xlab = "Residuals", col = "skyblue",
breaks = 20)

# QQ plot for Normality
qqnorm(residuals_211)
qqline(residuals_211, col = "red")

# Ljung-Box Test for White noise
Box.test(residuals_211, lag = 20, type = "Ljung-Box")

# Forecast next 40 periods using ARIMA(2,1,1)
forecast_arima <- forecast::forecast(model_211, h = 40)

# Plot the forecast
autoplot(forecast_arima) +
  ggtitle("Forecast - NY Metro Job Posting Index (ARIMA 2,1,1)") +
  ylab("Job Posting Index") + xlab("Date") + theme_minimal()
```