Portfolio: Time Series Model: Forecasting Air Quality Index in Mission, TX

Marissa Llamas

2024-11-18

Model Summary:

In this time series portion of the portfolio, we investigate methods to forecast the air quality index (AQI) in the Mission, TX site over the year 2022 and analyze any trends of the metric given data from previous dates in the year. We build an ARIMA(1,1,1) model that uses past AQI values from 2022, compiled by the U.S. Environmental Protection Agency, to compute these forecasts using the `forecast` package, yielding a low AIC value of 2301.61 among other tested models and no significant auto correlations at the first 50 lags and beyond. Using R, this model was constructed using Conditional Sum of Squares to find starting values then Maximum Likelihood Estimation to estimate model parameters, as documented by the ARIMA function, and we produce the following first-order auto regressive and first-order moving average model:

$$\triangle\hat{y}_t = -0.0594 + \hat{\epsilon}_t + -0.9323 \cdot \hat{\epsilon}_{t-1} + 0.4592 \cdot \triangle\hat{y}_{t-1}$$

$$\hat{y}_t - \hat{y}_{t-1} = -0.0594 + \hat{\epsilon}_t + -0.9323 \cdot \hat{\epsilon}_{t-1} + 0.4592 \cdot (\hat{y}_{t-1} - \hat{y}_{t-2})$$

$$\hat{y}_t = -0.0594 + \hat{\epsilon}_t + -0.9323 \cdot \hat{\epsilon}_{t-1} + 1.4592 \cdot \hat{y}_{t-1} - 0.4592 \cdot \hat{y}_{t-2}$$

Our model attempts to provide accurate forecasts for the daily AQI metric in Mission, TX from 2022, which can be used to inform public health decisions and enhance awareness about pollution and air quality.

Model Construction and Assessment:

We expect some form of seasonality in daily recorded AQI values, especially since this is a year long data set from 2022 and the AQI metric is dependent upon combinations of high temperatures, winds and breezes, and air borne particles. Moreover, the National Institutes of Health state, ozone, a pollutant, is produced with sunlight and high temperatures or stagnant air. That is, we can expect some slightly higher, hence unhealthier, AQI values in the summer months. However, the NIH also states this is often negligible with other major factors that determine AQI, like PM 2.5 as we've explored in our Linear Regression Model. In all, AQI does not necessarily have a seasonal component, and we choose to not compute a seasonal ARIMA model. *Source

When plotting the time series data alone, the plot shows non-stationarity. There does not seem to be a constant mean nor a constant variance, and there are general patterns of increasing and decreasing AQI values throughout the year. This calls for a need for at least one regular differencing.

The ACF plot for the raw time series data showcases some potential seasonality behavior of the time series with dampening oscillation. There is also a slow linear decay present, further adding support to our decision

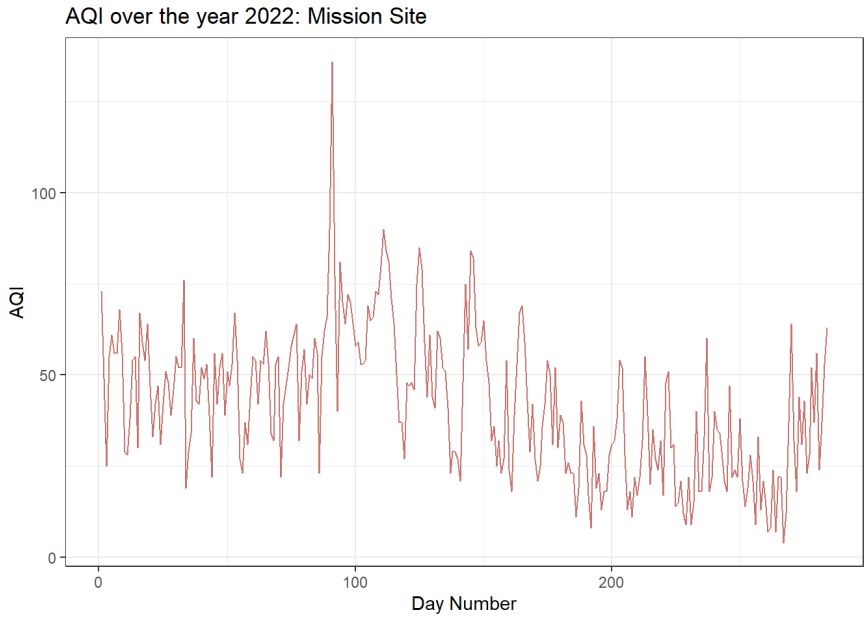

Figure 1: Time Series Plot for AQI in 2022 Mission, TX Site

to use at least some sort of differencing. However, we need to make our time series stationary first before we decide on a model.

When computing one regular difference on the time series data, the new time series plot shows stationarity with a constant mean centered at 0 and constant variance with no general increasing or decreasing patterns. Additionally, there doesn't seem to be clear, strong seasonality in the data since there are no spikes at regular intervals in the plot. Hence, seasonal differencing does not appear necessary and it is sufficient to use first differencing to make the series stationary.

To adjust if we can make this plot have a stronger, more present constant mean with constant variance, we attempt to do a second differencing on the time series data. However, there is no change in the behavior of the plot. Hence, we stick with just one regular differencing to keep the model simple.

Therefore, our ARIMA model must compromise of a differencing parameter of 1 and be of the form ARIMA(p, 1, q). To find our remaining ARIMA parameters, p, q, we construct both ACF and PACF plots for the first differenced time series. Displaying the ACF plot, we note there are two significant autocorrelations at lags 1 and 2, signifying potentially using second-order moving average terms to forecast AQI values.

Now, in the PACF plot, notably, there are four significant auto correlations at lags 1, 2, 6, and 8. However, taking into account the scaling of the y-axis, we need our underlying model to be simple, so we make a choice to focus on the first two significant autocorrelations. This is to say we will focus on potential second-order autoregressive terms to forecast AQI values.

We experiment with different combinations of these ARIMA parameters and we select the best model using AIC and the ACF and PACF plots of the model residuals as our criterions. As a result, the best model turns out to be ARIMA(1,1,1), with significant coefficients.

In Table 1, we record the resulting AIC, Log-Likelihood, and σ^2 values of each ARIMA Model tested that yield significant coefficients when we performed a z-test on them.

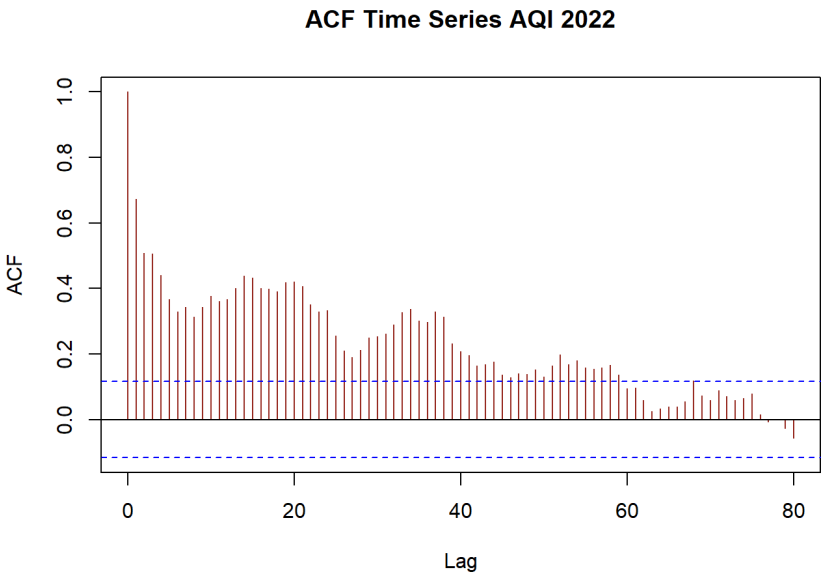

Figure 2: ACF Time Series Plot for AQI in 2022 Mission, TX Site

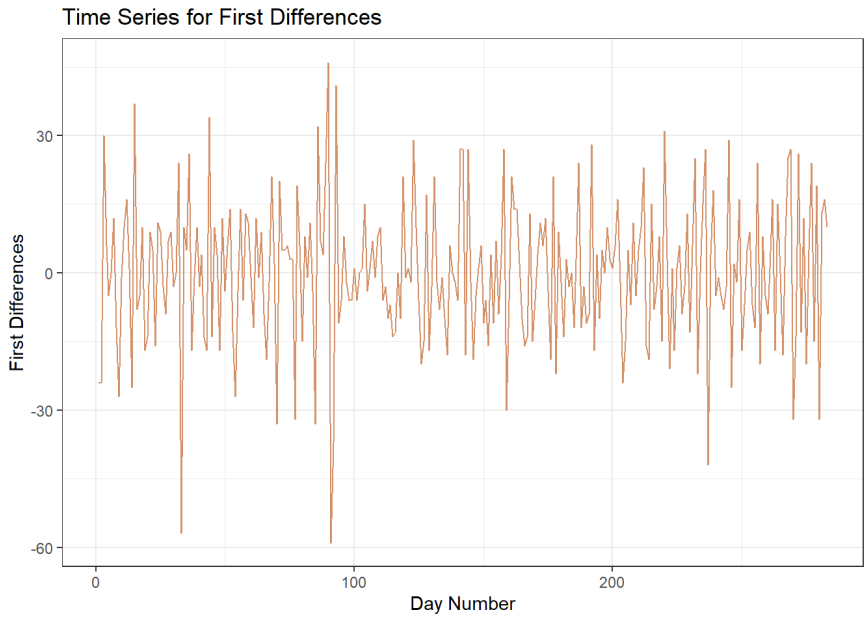

Figure 3: First Differences Time Series Plot for AQI in 2022 Mission, TX Site

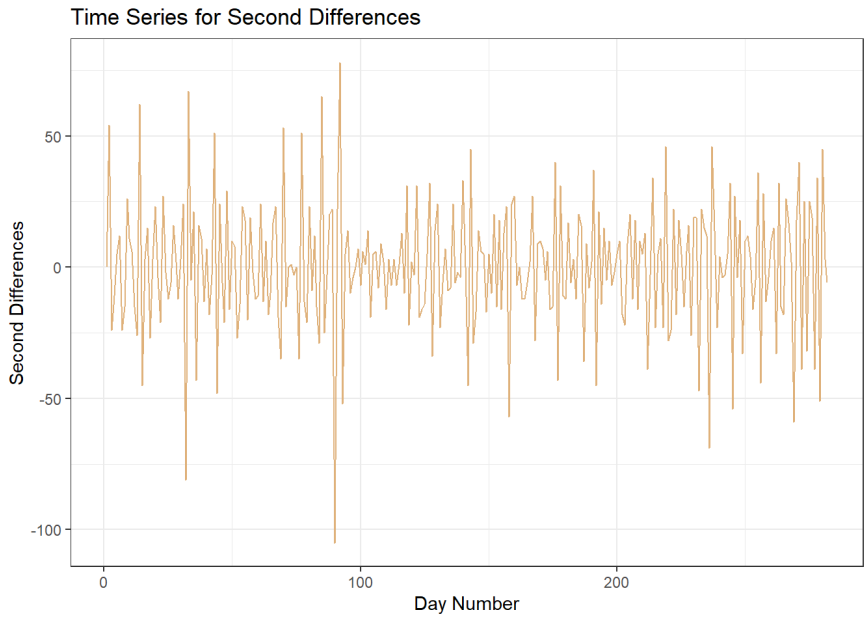

Figure 4: Second Differences Time Series Plot for AQI in 2022 Mission, TX Site

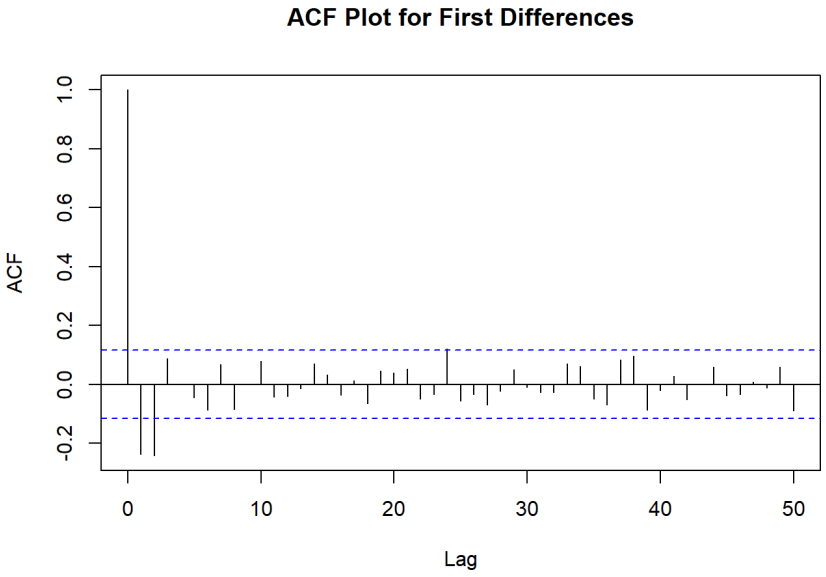

Figure 5: First Differences ACF Plot for AQI in 2022 Mission, TX Site

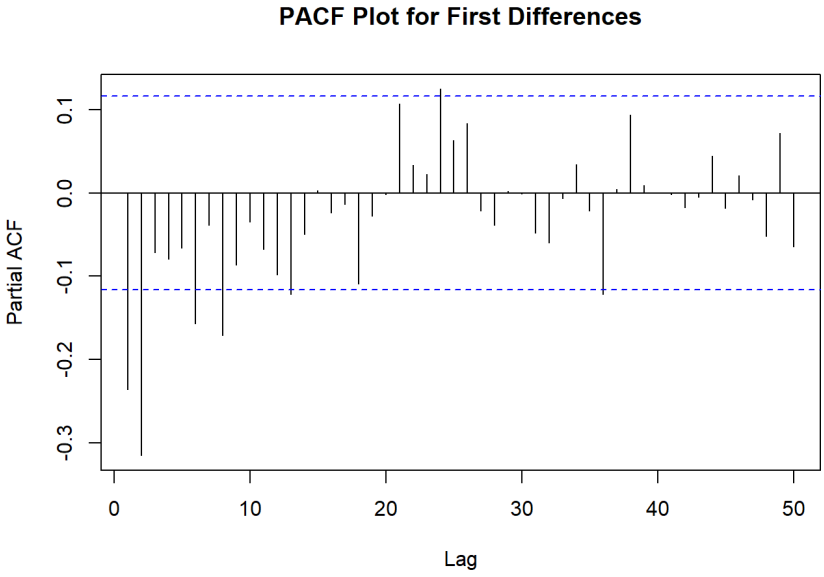

Figure 6: First Differences PACF Plot for AQI in 2022 Mission, TX Site

Table 1: ARIMA Model Selection

Model Type	AIC	sigma^2	Log Likelihood
(0, 1, 2)	2301.28	194.8	-1146.64
(2, 1, 0)	2322.45	210.6	-1157.22
(1, 1, 1)	2301.61	195.1	-1146.80
(0, 1, 1)	2330.17	217.2	-1162.09
(1, 1, 0)	2351.49	234.4	-1172.74

From these models, we can select between an ARIMA(1,1,1) and an ARIMA(0,1,2) since both have the lowest AIC and σ^2 values but only differ by a small amount. When we construct ACF and PACF plots for both models, however, the ARIMA(1,1,1) shows no significant autocorrelations at any lag values in the ACF plot, while the ARIMA(0,1,2) model has two significant autocorrelations at lags 6 and 8. This suggests the moving average terms of the ARIMA(1,1,1) model work well to forecast future values. In terms of the PACF plot, the ARIMA(1,1,1) and ARIMA(0,1,2) models have significant autocorrelations at lags 3 and 21, and lag 6 respectively. This signifies we need some improvement in the auto regressive terms. However, we make a choice to not complicate the model any further since the autoregressive terms we tested earlier either did not yield significant coefficients or had high AIC values.

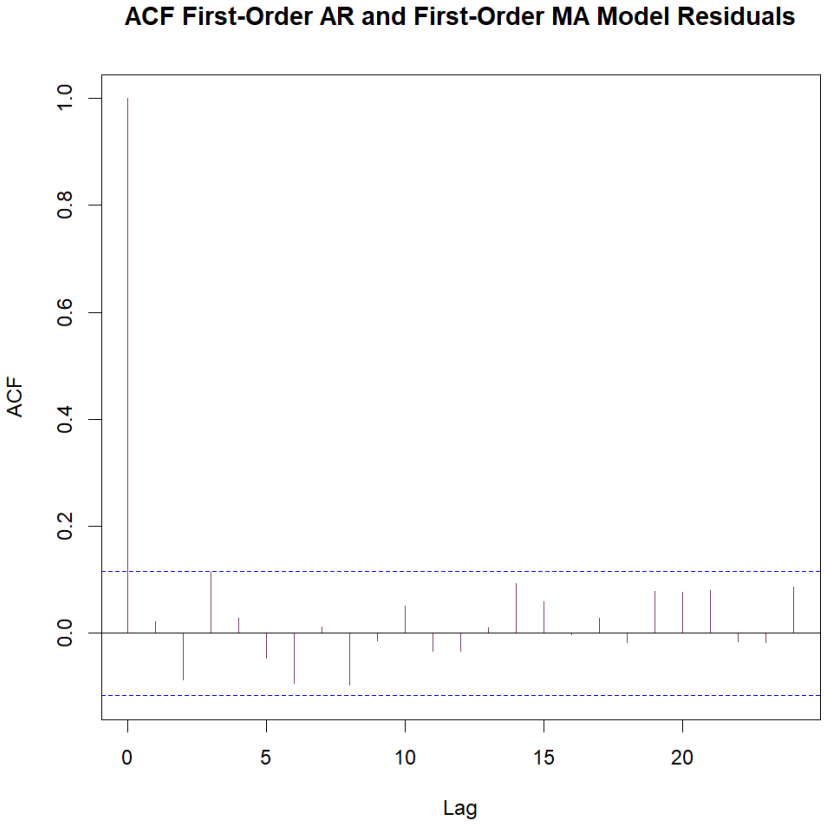

Figure 7: ACF for ARIMA(1,1,1) Residuals

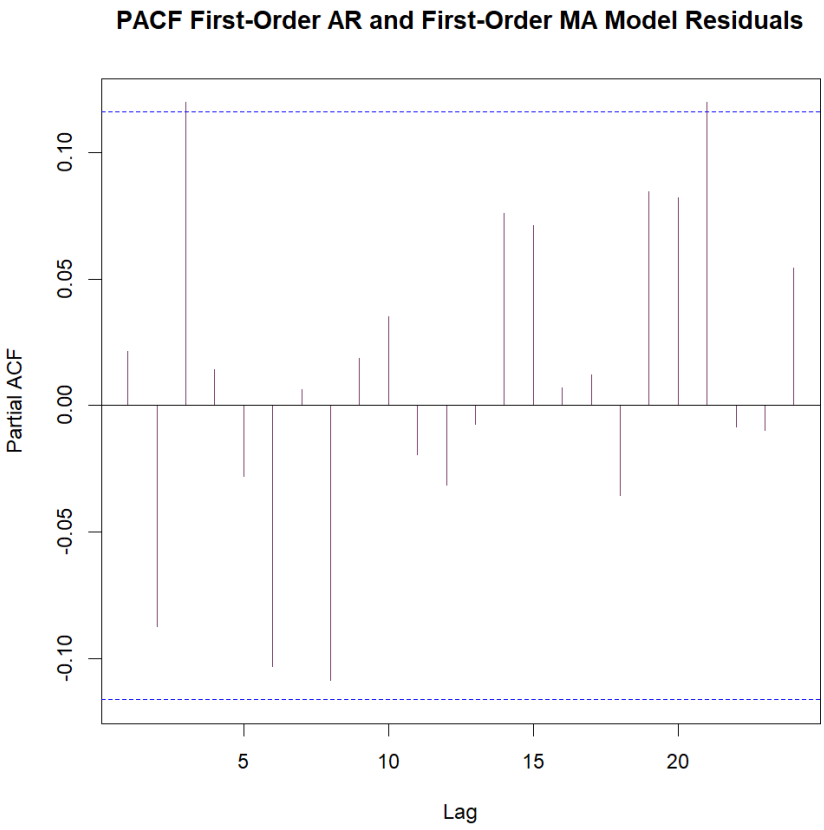

Figure 8: PACF for ARIMA(1,1,1) Residuals

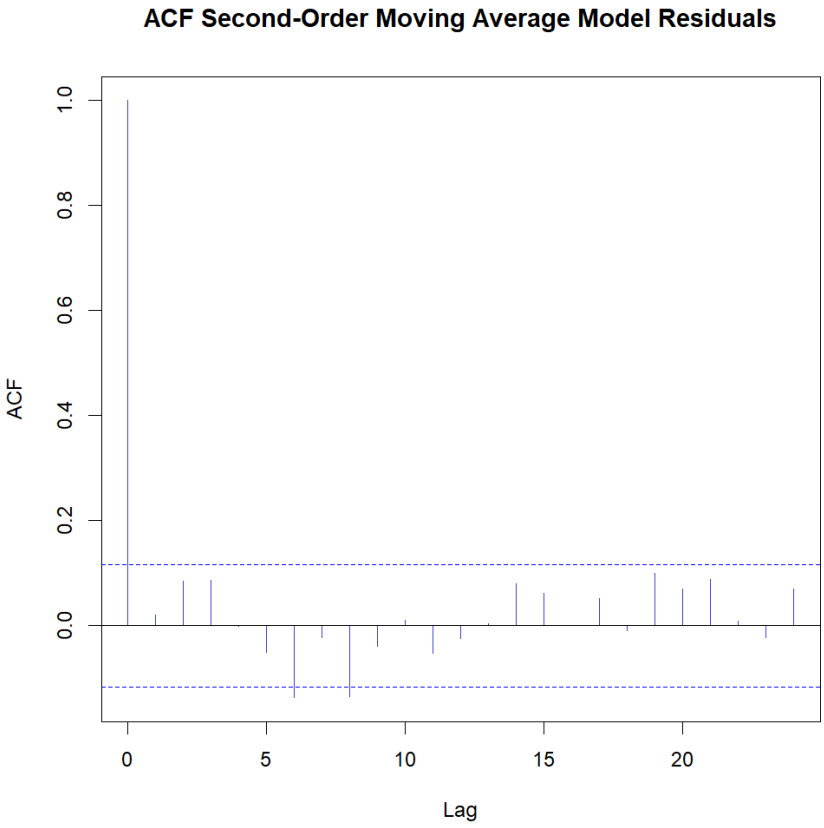

Figure 9: ACF for ARIMA(0,1,2) Residuals

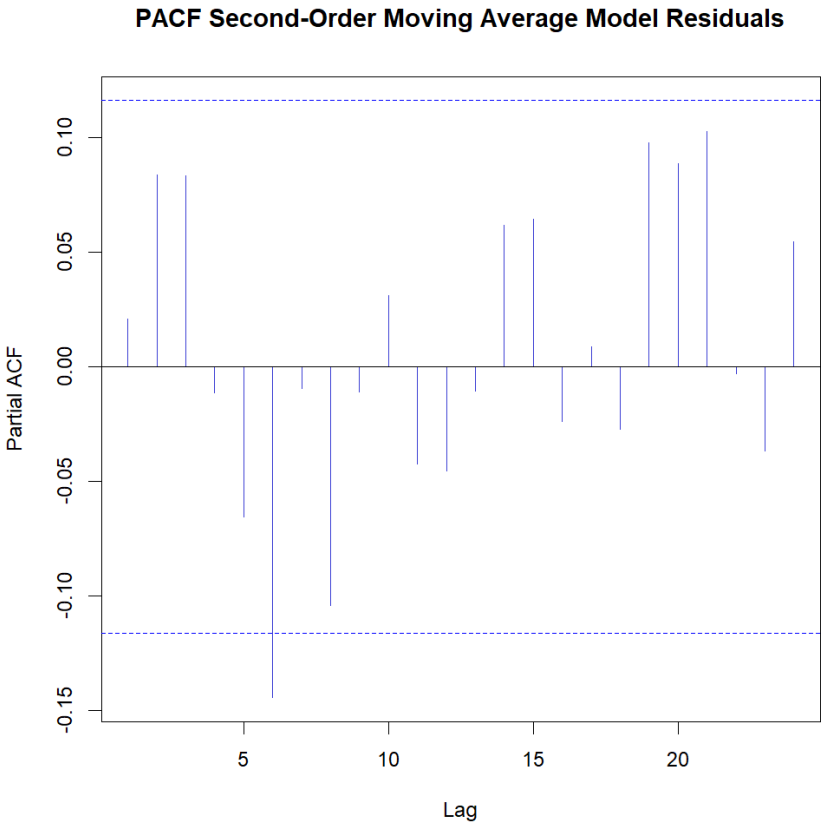

Figure 10: PACF for ARIMA(0,1,2) Residuals

So, we make the decision to choose the model with no significant autocorrelation residuals in the ACF plot at the cost of a slightly higher AIC and σ^2 values. As a result, the best model here is the ARIMA(1,1,1).

Below, we have the model residuals for the ARIMA(1,1,1) model and note that these residuals seem to follow a constant mean centered at zero and are normally distributed, all of which allows us to further validate and support our decision to proceed with this model.

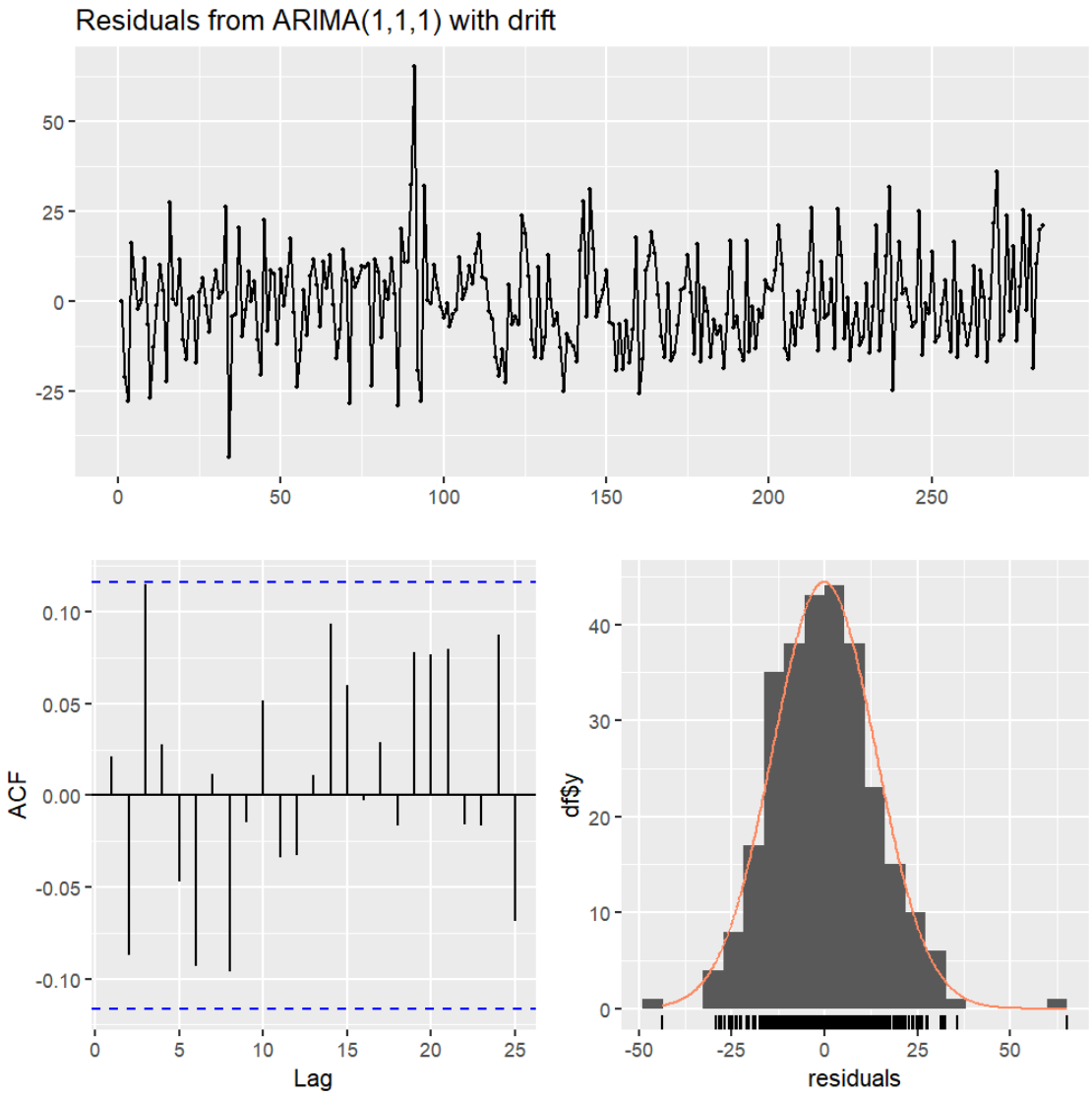

Figure 11: ARIMA(1,1,1) Model Residuals

The model produces coefficients for $\phi_1 = 0.4592$ and $\theta_1 = -0.9323$ with associated p-values of 3.478e^{-13} and $< 2.2\text{e}^{-16}$ respectively when conducting a z-test on each coefficient.

Model Usage and Concluding Thoughts:

This model can be used to forecast the next recorded daily AQI values in Mission, TX after our last recorded entry and produce prediction intervals for each estimate. Our last entry is December 13, 2022 with an AQI

of 63, so we can use our model to forecast the daily AQI value for the next 14 days. The table below displays the forecasted values and the prediction interval bounds for a 80% and 95% confidence level.

Table 2: 14-Recorded-Day Forecast for ARIMA(1,1,1)

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
285	47.92111	30.02232	65.81990	20.5472696	75.29495
286	40.96428	20.73268	61.19587	10.0227242	71.90583
287	37.73736	16.76022	58.71451	5.6555937	69.81913
288	36.22335	14.91241	57.53429	3.6310786	68.81563
289	35.49596	13.98359	57.00834	2.5956210	68.39631
290	35.12982	13.46560	56.79403	1.9972620	68.26237
291	34.92957	13.13420	56.72494	1.5964237	68.26271
292	34.80550	12.88826	56.72274	1.2859789	68.32502
293	34.71642	12.68178	56.75106	1.0173475	68.41549
294	34.64341	12.49370	56.79311	0.7683521	68.51846
295	34.57777	12.31437	56.84118	0.5288340	68.62671
296	34.51553	12.13936	56.89169	0.2941300	68.73692
297	34.45484	11.96663	56.94304	0.0620943	68.84758
298	34.39486	11.79525	56.99447	-0.1682643	68.95799

The point forecast for December 14th, 2022 is approximately an AQI value of 48 and the 95% prediction interval is $(20.54, 75.29)$. We can interpret this to say that we are 95% confident the true AQI value in Mission, TX for December 14th, 2022 will be between 20.54 and 75.29. This interval is particularly wide, so we can also make a narrower interval at the cost of a smaller prediction level of 80% to say that we are 80% confident that the true AQI value in Mission, TX for December 14th, 2022 will be in between $(30.02, 65.8)$. The point forecast is less than the recorded AQI value of the previous day (December 13th) so we can say we expect better air quality in December 14th than we experienced in December 13th, as the model forecasts a 23.8% decrease in AQI.

There are a few limitations to the model as it does not have access to some missing daily AQI values. For instance, we have a total of 284 recorded data points for the daily 2022 AQI in Mission, TX, but there are 365 days in a year. While this could have been mitigated by using a different site location, most data I found did not have complete 365 data points. It is interesting how we have stable, continuous dates for the first three month chunk of the year and majority of missing values at the July and December months. This could be due to broken monitors and repairments or system malfunctions. Still, this model serves as a way to better understand the behavior of AQI throughout the 2022 year in Mission, TX, my hometown, and inform entities like local governments and citizens to spread awareness of the cleanliness of the air around them and establish methods to regulate air quality.

R Appendix:

‘

```
# Loading libraries..
library(ggplot2)
library(forecast)
library(dplyr)
library(tidyverse)
library(knitr)
```

```r
# Data from U.S. Environmental Protection Agency
rgv_data = read.csv("C:\\Users\\mllam\\Downloads\\rgv_pm25_data_2022.csv")

# Cleaning Data: Renaming columns and filtering negative pm 2.5 values
mission_dat <- rgv_data |>
               filter(Daily.Mean.PM2.5.Concentration >= 0)|>
               rename(pm = Daily.Mean.PM2.5.Concentration)|>
               rename(aqi = Daily.AQI.Value) |>
               select(Date, pm, Units, aqi, County, Local.Site.Name)|>
               filter(Local.Site.Name == "Mission")

# Saving Time Series Data
ts_data <-mission_dat

# Visualizing Time Series
plot(ts_data$aqi)
```

```r
acf(ts_data$aqi)
```

```r
pacf(ts_data$aqi)
```

```r
# Taking First Differences
plot(diff(ts_data$aqi))
```

```r
acf(diff(ts_data$aqi))
```

```r
pacf(diff(ts_data$aqi))
```

```r
# Taking Second Differences
plot(diff(diff(ts_data$aqi)))
```

```r
# Building ARIMA(1,1,1) model
model = Arima(ts_data$aqi, c(1, 1, 1), include.constant = TRUE)

# Validating model through Z-tests on coefficients
require(lmtest)
coeftest(ar1_ma1)

# Checks Residuals of Model
checkresiduals(model, col = "#ff29b0")
```

```r
# Forecast Values
forecast(model, h = 14, fan = TRUE)

# Plots Forecasted values
autoplot(forecast(model, h = 14, fan = TRUE), fcol = "#ff96d5", flwd = 0.9 ) +
  labs(x = "Day Number", y = "AQI") +
  theme_bw()
```