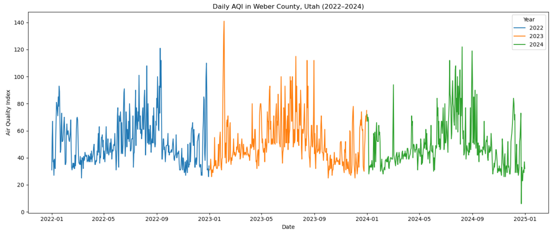
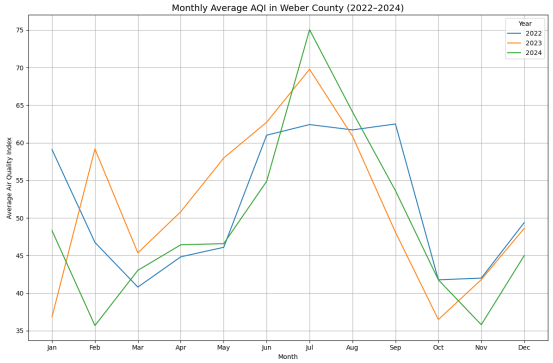
# Math6450 Assignment3: Time Series AQI Analysis

Ian Tai Ahn
October 20, 2025

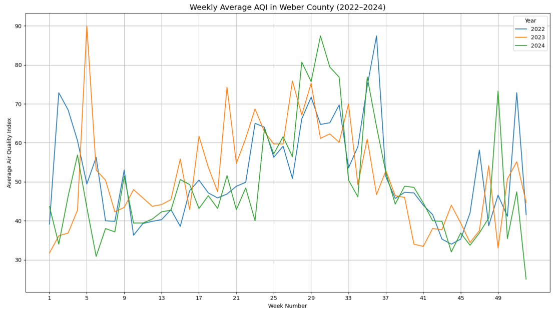
**Exploratory Data Analysis (EDA):** The data was loaded in and filtered to remove all columns except for State Name, and county Name. Then additional steps were taken to only grab rows with Utah, and Weber in their state and county name. 2022 had 365 rows, 2023 had 365 rows, and 2024 had 366 rows. This meant that there was a valid data entry for every single day, and this was proven to be true since there were no missing or NA values in these datasets.



The daily AQI shows that there is quite a bit of fluctuations day to day, and that each year follows a similar pattern. This means there is seasonality is present in this data.



Using the monthly average of the AQI shows a smoother line and a bigger picture view of how the data has changed over the years. However, it can smooth the line too much and it may not be able to capture short-term AQI fluctuations and changes. Smoothing out the graph also shows the obvious point where we have a seasonal trend. The month of July consistently jumps up in AQI.

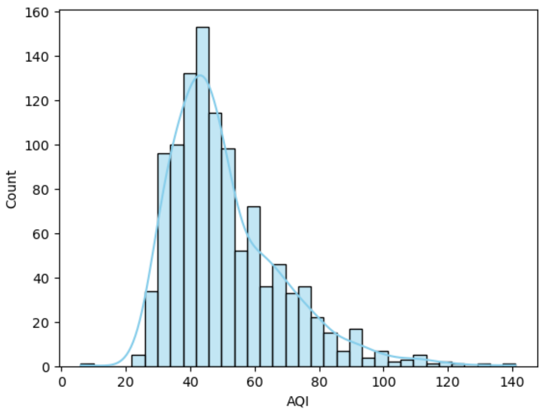


Looking at the weekly changes in AQI shows that there seems to be some seasonality and trends in the data, but we can see short-term fluctuations as well. This may be an applicable middle ground since the lines aren't as smooth as they are in the monthly AQI plot.

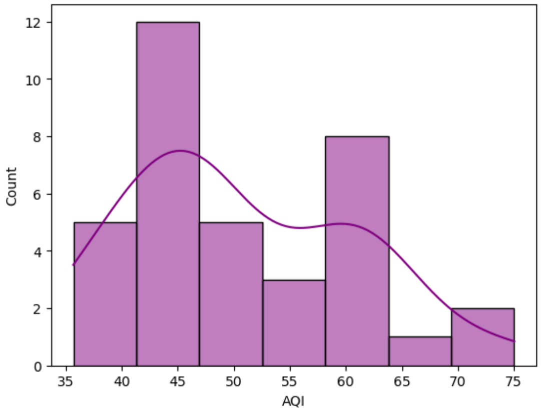
**Reason for Aggregation:** Upon initial data exploration, while there is enough data for daily time series modeling, the distribution was skewed right, and the daily trends were quite jagged. So, when graphing both monthly and weekly averages the monthly data looked a little too smooth, and weekly data looked to be the most promising since there are still obvious seasonal patterns and the data didn't look so smooth over it wouldn't capture other trends.

So, to capture as many AQI patterns as possible the weekly and monthly data will be used to compare results.

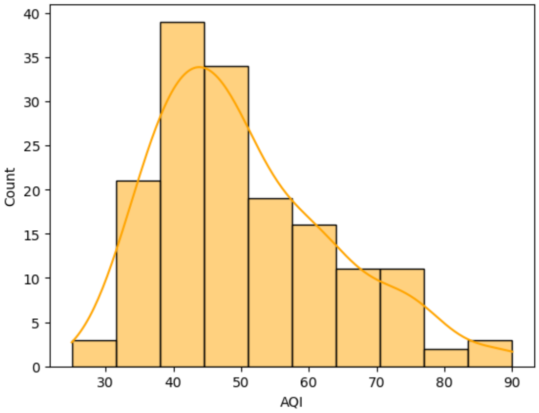
This means that from the daily AQI data we will aggregate the weekly and monthly means for analyzing, and AQI forecasting.



The daily AQI variance is a little skewed to the right, and so aggregation will be a good idea for making the variance more normal.

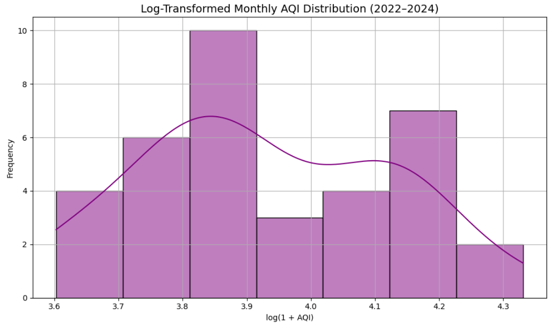


The monthly AQI data shows a similar right-skewed pattern, but the tail is a little smaller now.

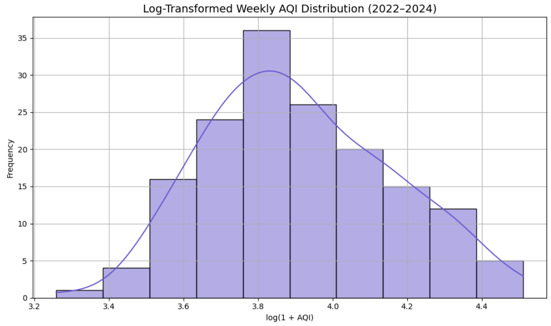


The variance is not as prevalent, but still there, so we will move to log transformations to ensure the data is stationary.

**Transformations:**

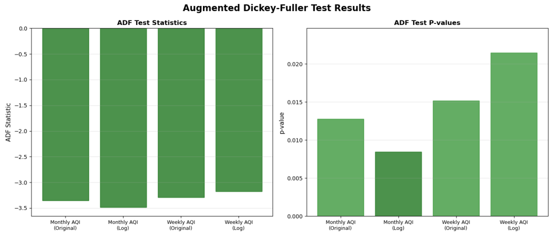


The distribution looks quite normal now after log transformation.

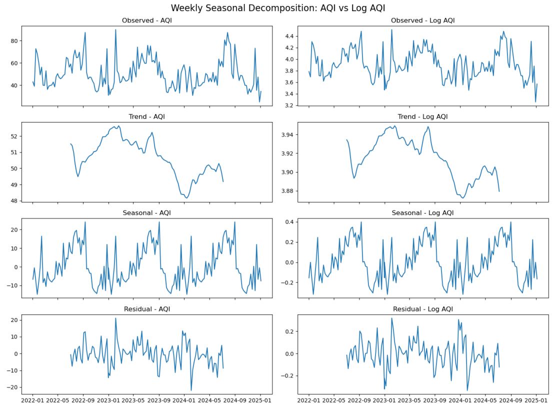


The log transformed weekly data also shows a more normal variance when plotted.

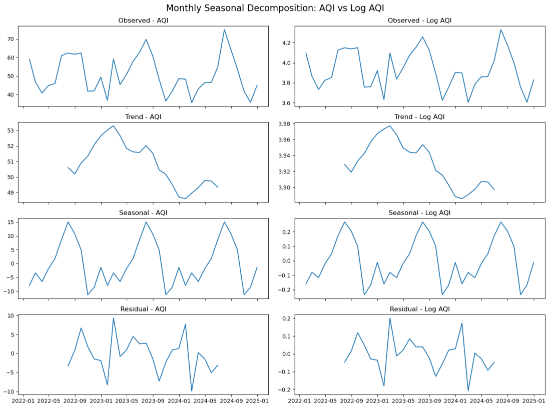
**Seasonal Patterns:** We first noticed the seasonality of the data in the Daily AQI graph, and it was made clear there was seasonality when graphing the aggregated monthly avg AQI dataset. To get rid of this seasonality seasonal differencing will be needed to ensure this data set is stationary and ready for modeling.



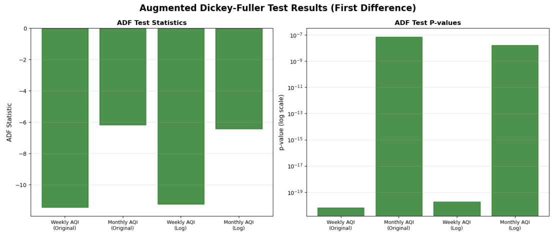
We can interpret the adf statistic as the more negative, the more evidence there is that the data is stationary. A smaller p-value means that there is high evidence we can reject the null resulting in a stationary time series. Both numbers are low so we can assume we have stationary time series data.



The seasonal decomposition shows there are several spikes in data here and there, but most importantly there is oscillating behavior in the visuals which means air quality does suffer from seasonality.



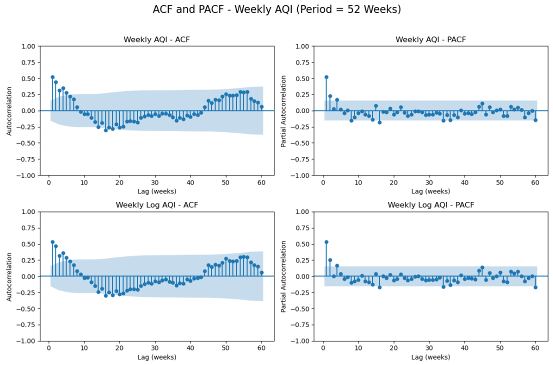
Since the trend and seasonality of this data was so obvious when looking at the values graphed, I wanted to also see what the monthly data looked like, and it is much more obvious that there is a major trend going on in this data.



We knew that there was seasonality in the data so by taking the season first difference we have adf stats that are more negative, and p-values that are even smaller, showing that the data has significant evidence that it is stationary.

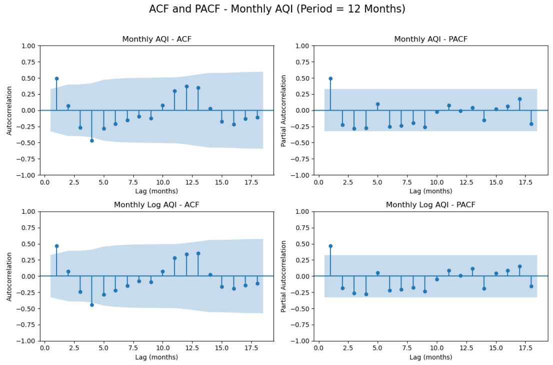
Comparing the ADF p-value for the weekly aqi/logged-aqi values and the monthly aqi/logged-aqi values show that the p-value is very small meaning were making the data more stationarity.

**Autocorrelation:**



The lags show that at about lag 8, we start to see a reverse of direction in correlation. This must mean that as it looks back further in the year, since the weather changes this also changes AQI.

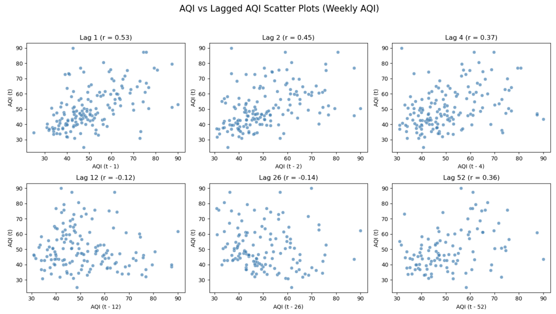
The PACF graph starts high, and after 1 lag has a steep drop off and then tails off with no clear pattern. This can be interpreted as after 1 lag, there isn't a direct strong influence on the weekly AQI values.

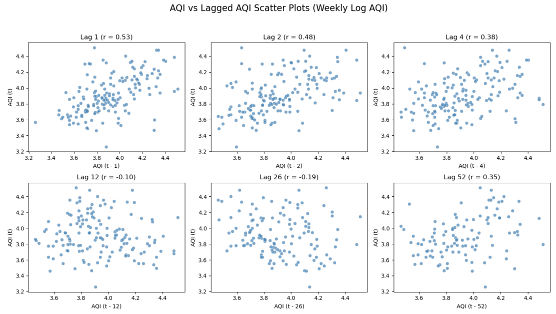


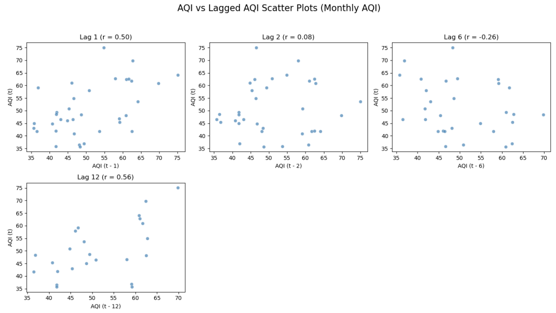
The monthly ACF graphs show a similar pattern to the weekly ACF graphs. There is usually a seasonal pattern that can be explained by the changing of the seasons showing that there is a correlation effect on AQI from the seasonal changes.

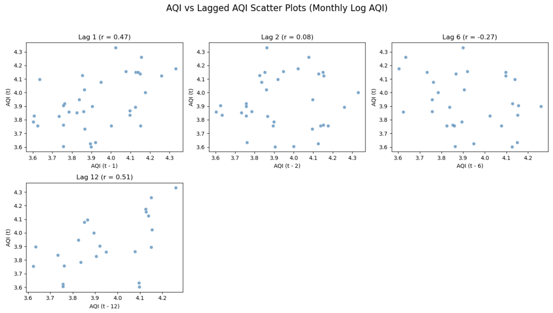
The monthly PACF graph also shows similar results to the weekly PACF plot results because of the immediate steep drop off on lag 1, and then there is no clear pattern. This confirms the verdict above that there is no strong and direct influence from earlier months once the 1st month is accounted for.

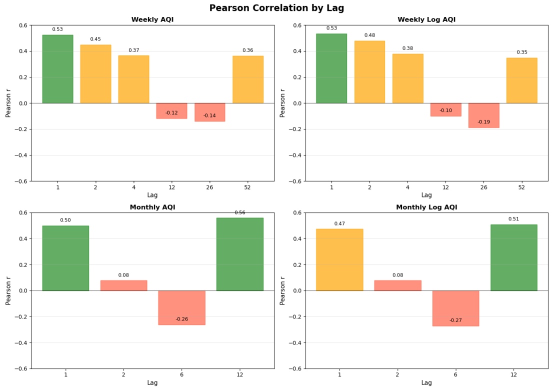
## Correlation Coefficients:









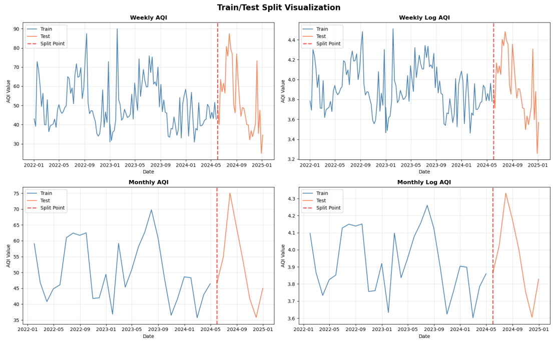


The weekly correlation shows seasonality due to what I'd assume literal seasons, like fall, winter, summer, and spring. It makes sense that on a lag of 1, there is a positive correlation because it is still within the season. However, looking at lags 12, and 26, that is comparing data to a different season which causes a negative correlation reversing the correlation. I can say since I started this lag in January, 26 weeks ago January is in the summer, June or July, so it makes sense the lag would calculate a negative correlation.

Looking at the monthly data pretty much confirms this as well, since looking at lag 6 shows a negative correlation with the current value meaning winter is colder than summer. So, we can confirm that there are correlation and seasonality in this data, and we will now be able to select a model to forecast the AQI.

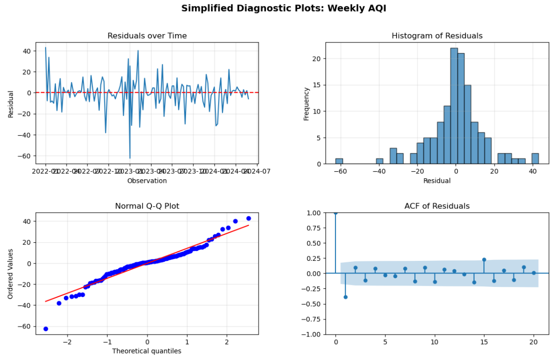
Comparing the pearson_r scores with the plotted ACF graph shows they both agree with each other reinforcing that correlation and seasonality is present in this data due to changing seasons.

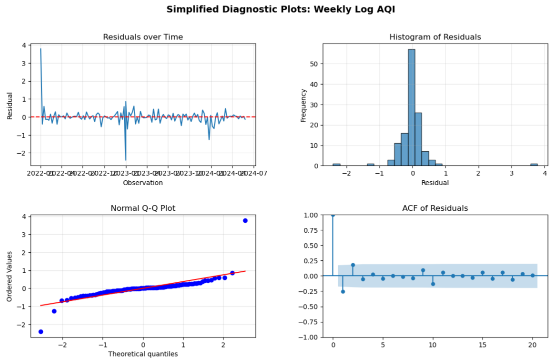
## Data Splitting:

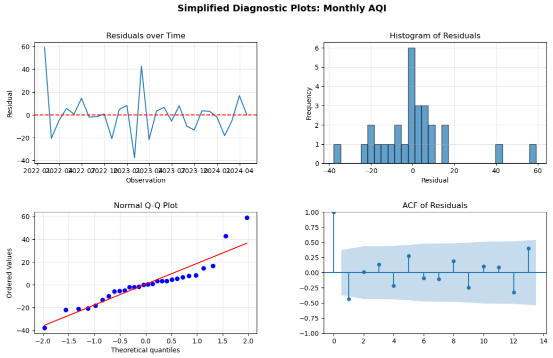


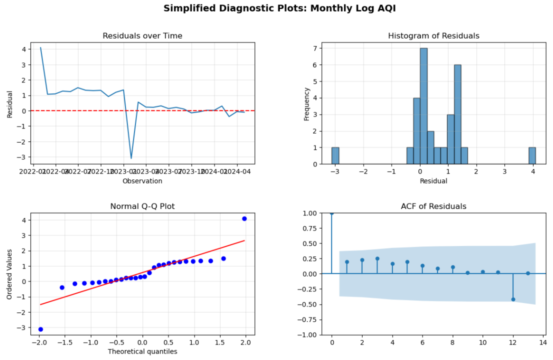
**Model Selection:** I tested different p, d, q, P, D, and Q values for the SARIMAX model, and made decisions based on the ACF/PACF plots and how stationary my data was. What I found was that taking the difference and seasonal difference impacted all the models except for the logged aqi positively. Also, setting p/P and q/Q to a value of 1 also showed better forecasting and results. These values set to 1 helped the model capture short-term autocorrelation which we saw a little bit of in the ACF plots, and the seasonal trends were apparent from graphed data so both p/P and q/Q helped the model perform well.

## Model Parameters and Diagnostics:

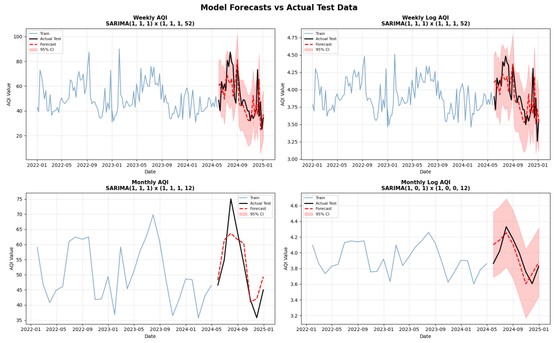




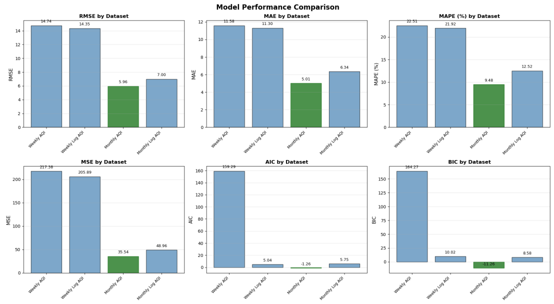




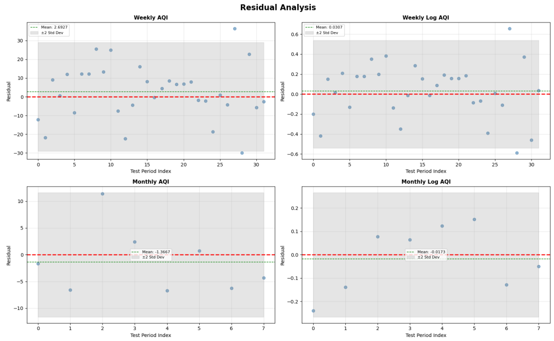
## Forecasting:



## Model Performance:



## Residual Analysis



## Final Model Equation:

```
Monthly AQI:
  SARIMA(1, 1, 1) x (1, 1, 1, 12)
  Notation: SARIMA(p=1, d=1, q=1) x␣
 ↪(P=1, D=1, Q=1, s=12)
```

```
  Model: (1 -  B)(1 - Φ B^12)(1 - B)(1␣
 ↪- B^12)Y  = (1 +  B)(1 + Θ B^12)

Best results by different criteria:
  Lowest RMSE: Monthly AQI (RMSE: 5.
 ↪9618)
  Lowest MAE: Monthly AQI (MAE: 5.0084)
  Lowest MAPE: Monthly AQI (MAPE: 9.
 ↪48%)
  Lowest AIC: Monthly AQI (AIC: -1.26)
```

**Conclusion:** This was an interesting study for many reasons. Firstly, since the time series data did have seasonality, it was a good exercise to discover what order difference, and seasonal differences to apply to the data. Since the dataset was populated quite well, and there weren't any missing dates, there was no need for dropping N/A values or imputing so that was nice to not have to handle that. Focusing on the ACF/PACF, stationarity, and variance let me think more about the seasonality of the data, and which p/d/q values to choose for the arima model.

In the end the best performing model was the base monthly AQI model that was aggregated monthly by average. However, the forecasted results given by this best performing model still couldn't capture the spike, similarly to most of the other models. I wonder what could have been done to the data/modeling that would capture the spike for the last couple weeks, or months in the dataset.