**OPIM 5671: Time Series Forecasting**

**Coffee Price Dynamics in the Face of Changing Brazilian Weather**

Lakshmi Ravi Chandu Kolusu

MS Business Analytics and Project Management

Specialization: Data Science

**Introduction:**

The coffee industry holds immense global importance, and Brazil plays a central role as the world's top coffee producer, contributing approximately one-third of the global coffee supply. In recent years, climate change has made weather patterns more unpredictable, causing increased uncertainty in coffee production. Given that Brazil's coffee production is strongly influenced by weather conditions, especially droughts and frosts, it has become crucial to understand how changes in Brazil's weather, specifically temperature and precipitation, impact coffee production and prices. Our project aims to investigate this vital connection to provide valuable insights that can help coffee producers, traders, and other stakeholders effectively manage risks within the coffee market.

The primary objective of this project is to develop a forecasting model for coffee prices based on variations in temperature and precipitation within Brazil's coffee-growing regions. By analyzing historical weather data alongside coffee price movements, this project aims to identify key patterns, correlations, and trends. The ultimate goal is to provide a valuable forecasting tool that can assist coffee industry stakeholders in anticipating price fluctuations driven by temperature and precipitation changes. A limitation to this project is that we are not able to analyze every region in Brazil. We chose to focus on only the Southeast region since that is where the vast majority of coffee is grown in Brazil. Even though we had a specific region, we chose to focus on the state of Minas Gerais since there is substantial coffee production in this region, producing approximately half of Brazil's coffee, and it was given in our chosen dataset.

The research question that guides this project is: Can a time series forecasting model accurately predict coffee prices based on temperature and precipitation variations in Brazil's most prominent coffee growing regions? To address this question, the study will explore various time series modeling techniques, including traditional statistical methods like ARIMAX and SARIMAX, as well as deep learning models. By comparing the performance of these models and identifying the most effective approach, the project seeks to make a valuable contribution to the ongoing efforts to predictcoffee price dynamics in the face of changing Brazilian weather.

**Data Description:**

**Raw Data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Source** | **Reference Link** | **Rows** | **Columns** |
| [Climate Weather Surface of Brazil](https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region) | Kaggle | <https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region> | 15,345,216 | 27 |
| [Coffee Price Historical Data](https://www.kaggle.com/datasets/timmofeyy/coffee-prices-historical-data) | Kaggle | <https://www.kaggle.com/datasets/timmofeyy/coffee-prices-historical-data> | 12,564 | 2 |

**Combined Dataset:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Reference Link** | **Rows** | **Columns** |
| MonthlyData2013-19(entire\_weather\_features\_MG) | https://drive.google.com/file/d/1QB1lQSFbOSGVtAP01C-LhFJNZDH7EFNM/view?usp=drive\_link | 85 | 7 |
| MonthlyData2013-19(entire\_weather\_features\_Patrocinio) | https://drive.google.com/file/d/1-9WY2EqkmB1WKiwkXREeBYarW3lDWKYJ/view?usp=drive\_link | 85 | 7 |

**Overview of Variables:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Variable Name** | **Description** | **Type** |
| Target Variable | Mean\_Price | Mean Coffee Price in USD per Pound | Numerical |
| Input Variable | Mean\_Air\_Temp | Monthly Minas Gerais Mean Air Temperature (°c) | Numerical |
| Input Variable | Min\_Air\_Temp | Monthly Minas Gerais Minimum Air Temperature (°c) | Numerical |
| Input Variable | Max\_Air\_Temp | Monthly Minas Gerais MaximumAir Temperature (°c) | Numerical |
| Input Variable | Mean\_Precipitation | Monthly Minas Gerais Mean Precipitation (mm) | Numerical |

**Data Exploration & Manipulation:**

[**Coffee Price Historical Data**](https://www.kaggle.com/datasets/timmofeyy/coffee-prices-historical-data)**:**

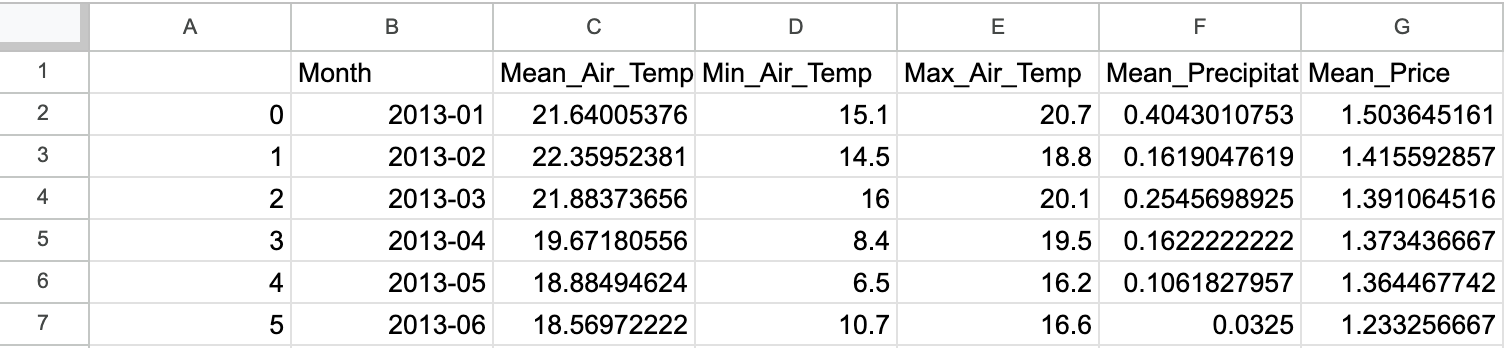
When exploring this dataset, we found that there were 21 null values ranging from May 8, 2023, to August 25, 2023. Based on our observations, we decided to drop these values since they were the most recent observations that had not been inputted yet. It was determined that these values had just not been recorded and were not within our selected time span (2013 to 2019). After subsetting the data to our desired time span of 2013-01-01 to 2019-12-31, we observed that there was a significant amount of missing dates. This was an issue since we needed our data to have equal intervals for dates. We inserted new rows for these missing dates and then forward-filled the values for each date. This ensured that the structure of the data was maintained while also ensuring there was no missing data.

[**Climate Weather Surface of Brazil**](https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region) **:**

The original source from this dataset had csv files for the different regions of Brazil: Central West, North, Northeast, South, and Southeast. For our purposes we decided to only use the Southeast data file since that is where the vast majority of coffee is grown in Brazil. We first translated and renamed columns from Portuguese to English. We then subsetted the data to only include the state of Minas Gerais. Within this subset, we removed all rows except for Air temperature (°c), Min. temperature (°c), Max. temperature (°c), and Precipitation (mm). We then filled in the missing values (or -9999) with the forward-fill method. Next, we aggregated Min temp by Min and Max temp by Max to find the absolute lowest and highest temperatures. This transformation gave us daily data instead of hourly data for each category.

**Combined Datasets:**

Once both of our datasets had been cleaned and formatted correctly, we were able to merge the data on the ‘date’ column. Since we had daily coffee price data and had transformed the Brazil climate data into daily data, this was the most effective way to create a new dataset. With the new combined data, we then grouped by the month of the date column for each category. This allowed us to have a dataset that had a row for each month of the year from 2013-01-01 to 2019-12-31. The columns in our final dataset are Month, Mean\_Air\_Temp, Min\_Air\_Temp, Max\_Air\_Temp, Mean\_Precipitation, and Mean\_Price (Exhibit 1). In these created feature columns the values are for example the minimum of minimum temperature for the month, the mean of the mean temperature for the month, etc. We transformed the hourly data into daily, calculated the minimum air temperature, then grouped by month and calculated the minimum air temperature for each month.

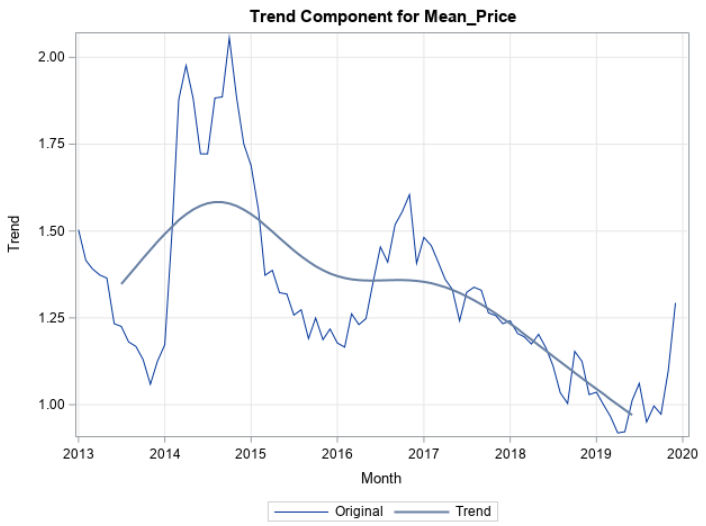
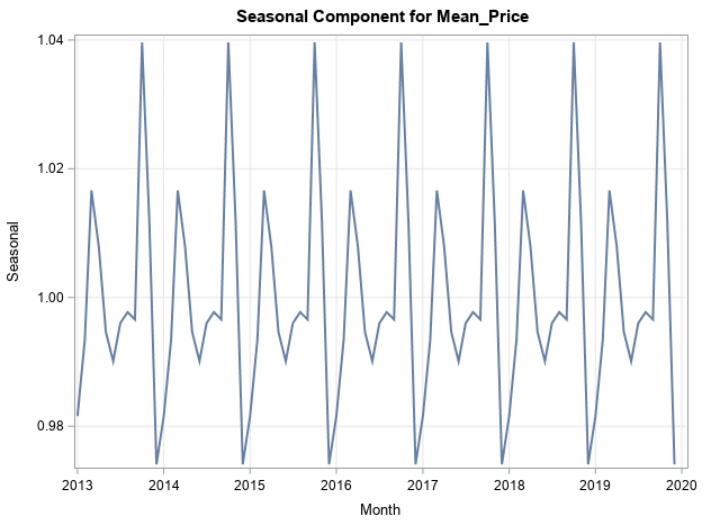
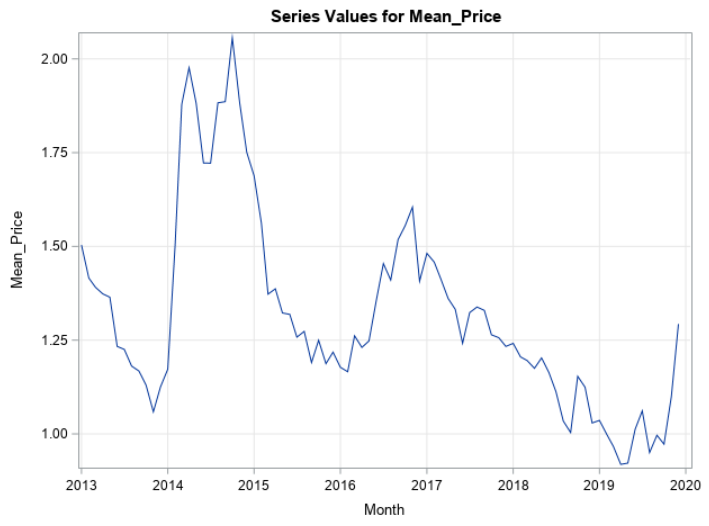
****

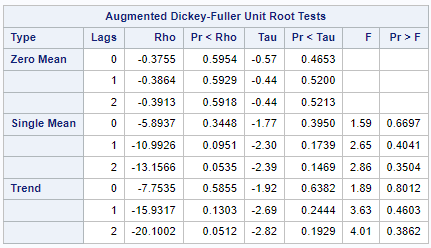
**Exhibit 1:** Screenshot of Finalized Dataset

**Parameter Analysis:**

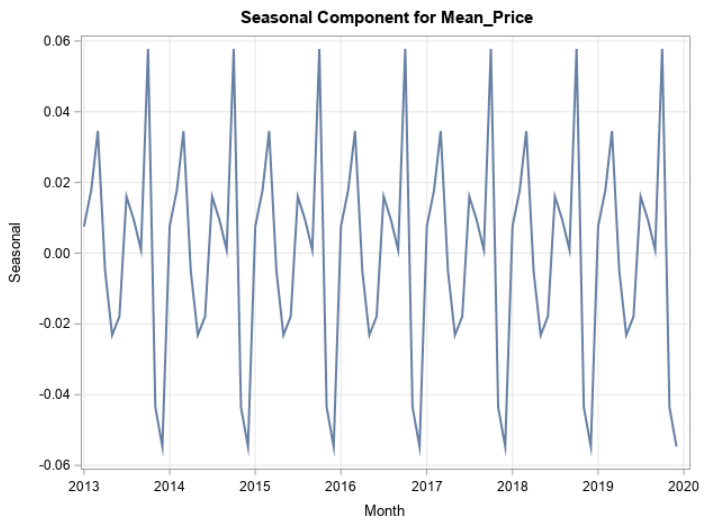
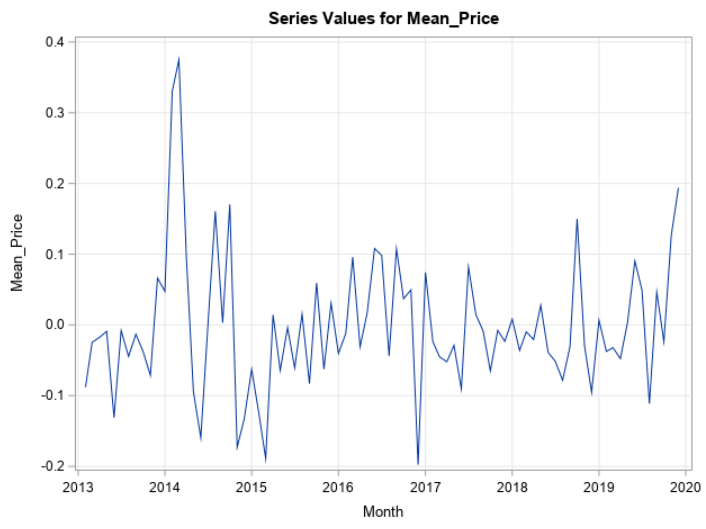
**Target Variable Analysis:**

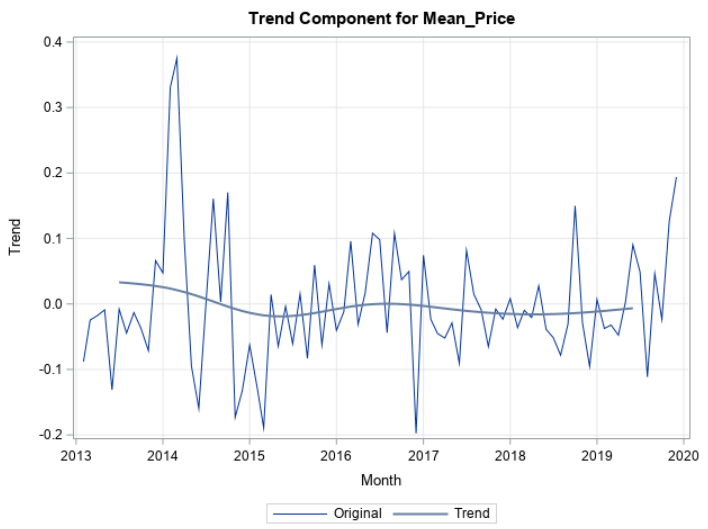
To gain a better understanding of our Target Variable ‘Mean\_Price’, we explored this variable and found it exhibited both seasonality and trend. We also conducted an Augmented Dickey-Fuller (ADF) unit root test. Our null hypothesis was that the data contains a unit root and therefore the data is non-stationary. Unfortunately, the tests returned p-values greater than our chosen significance level of 𝞪 = .05 for both the F-Statistic and our lagged difference Tau. From this, there is not enough statistical evidence to conclude our data is stationary and does not include a unit root.

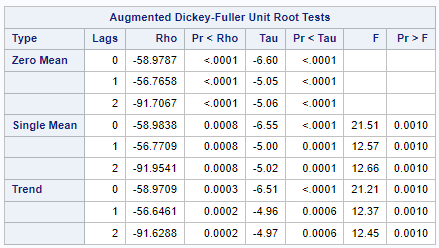




After concluding our data was not stationary, we conducted a first-order differentiation to attempt stationarity. When exploring the transformed data, we found that it exhibited seasonality but no trend. We again conducted an Augmented Dickey-Fuller (ADF) unit root test and found that our data is stationary. The results from this test show that the p-value for the F-Statistic and our lagged difference Tau are all less than .05.

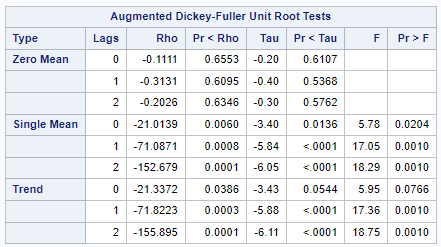
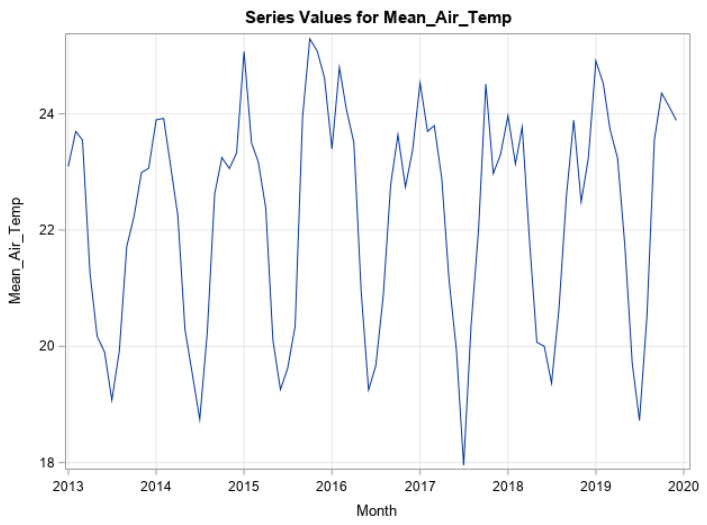


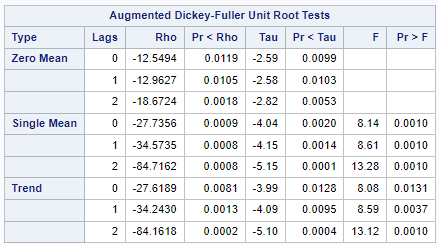
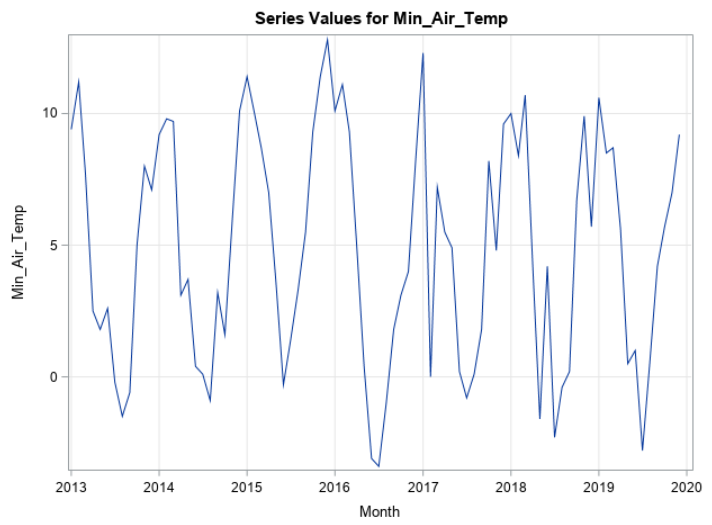


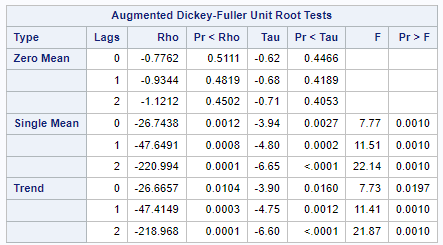
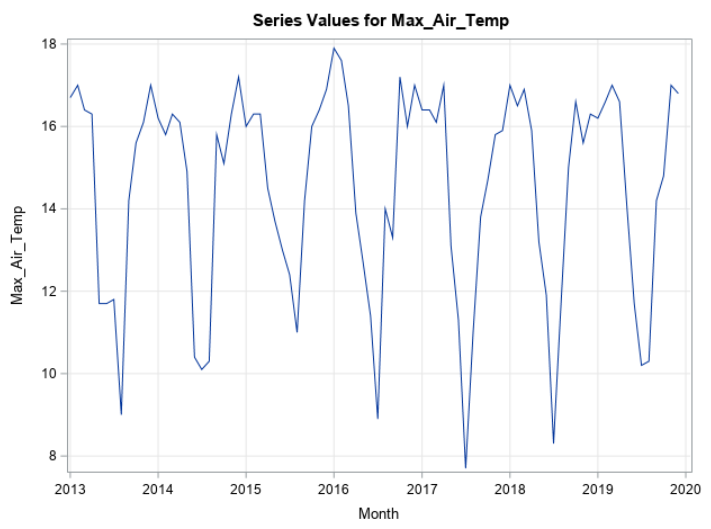


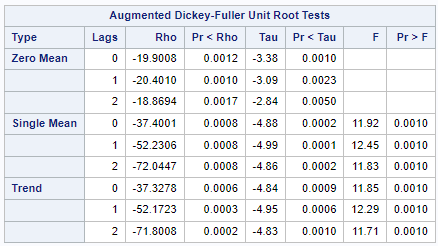
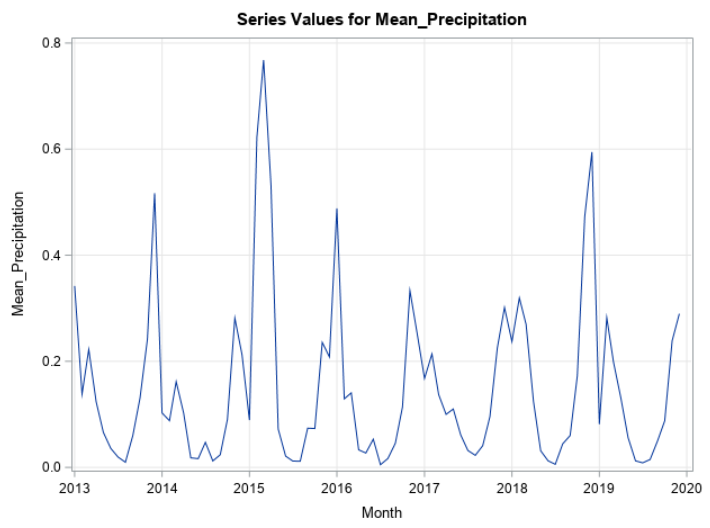
**Explanatory Variable Analysis:**

We next explored our input variables: Mean\_Air\_Temp, Min\_Air\_Temp, Max\_Air\_Temp, and Mean\_Precipication. For every input variable, there was no trend observed but there was strong seasonality exhibited. We then conducted an Augmented Dickey-Fuller (ADF) unit root test for each variable and found that our data is stationary. The results from this test show that the p-value for the F-Statistic and our lagged difference Tau are all less than .05.



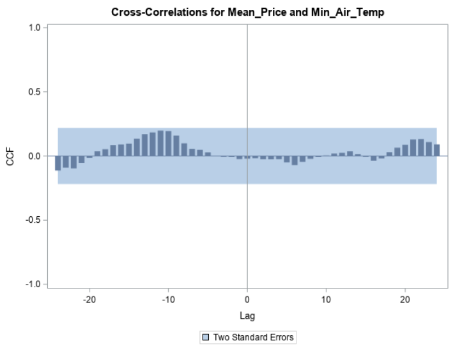
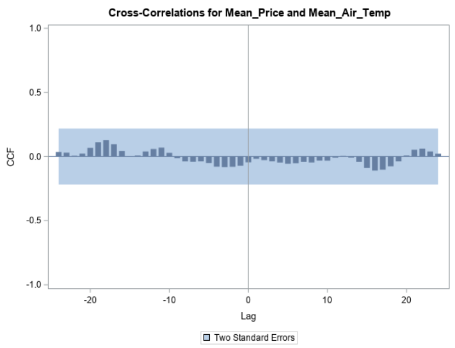


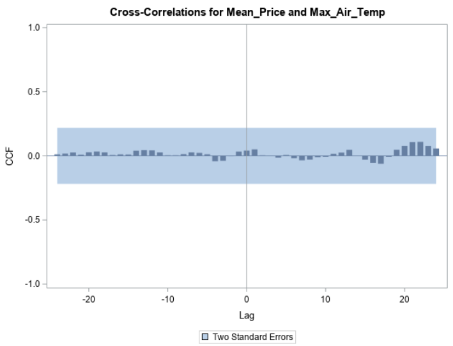




**Cross Correlation Analysis:**

After analyzing the behavior of our variables, we conducted cross correlation analysis on our target variable ‘Mean\_Price’ against all of our input variables individually. Before prewhitening, there was no lag effect observed from the cross-correlation plots.





**Prewhitening:**

We selected the best ARMA model to eliminate autocorrelation for all x variables based on the AIC, SBC criteria, and the presence of residual white noise. We then applied the same model individually to the y variable in combination with the best x models and examined its cross-correlation function (CCF) results.

1. Mean\_Air\_Temp

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| AR | MA | AR(S) | MA(S) | AIC | SBC | WN |
| 0 | 0 | 1 | 1 | 223.3001 | 230.5926 | No |
| 1 | 0 | 1 | 1 | 205.1692 | 214.8925 | Yes |
| 0 | 1 | 1 | 1 | 209.9308 | 219.654 | No |
| 1 | 1 | 1 | 1 | 207.169 | 219.3231 | Yes |

For Mean\_Air\_Temp, the model with AR(1), MA(0), seasonal AR(1), and seasonal MA(1) components was selected.

1. Min\_Air\_Temp

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| AR | MA | AR(S) | MA(S) | AIC | SBC | WN |
| 0 | 0 | 1 | 1 | 416.7512 | 424.0437 | No |
| 1 | 0 | 1 | 1 | 413.059 | 422.7822 | No |
| 0 | 1 | 1 | 1 | 414.2959 | 424.0192 | No |
| 1 | 1 | 1 | 1 | 414.0664 | 426.2205 | No |
| 2 | 0 | 1 | 1 | 414.2794 | 426.4335 | No |
| 2 | 1 | 1 | 1 | 416.0653 | 430.6502 | No |
| 0 | 2 | 1 | 1 | 415.1242 | 427.2783 | No |
| 1 | 2 | 1 | 1 | 415.4654 | 430.0503 | No |
| 2 | 2 | 1 | 1 | 430.8363 | 447.852 | No |

For Min\_Air\_Temp, the model with AR(1), MA(0), seasonal AR(1), and seasonal MA(1) components was selected. We selected the best option among those identified, which exhibited the lowest AIC and SBC values, as none of the ARMA models demonstrated a significant improvement over the white noise model.

1. Max\_Air\_Temp

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| AR | MA | AR(S) | MA(S) | AIC | SBC | WN |
| 0 | 0 | 1 | 1 | 283.8927 | 291.1852 | Yes |
| 1 | 0 | 1 | 1 | 285.849 | 295.5722 | No |
| 0 | 1 | 1 | 1 | 285.8456 | 295.5689 | No |
| 1 | 0 | 1 | 1 | 287.8002 | 299.9543 | No |

For Max\_Air\_Temp, the model with AR(0), MA(0), seasonal AR(1), and seasonal MA(1) components was selected.

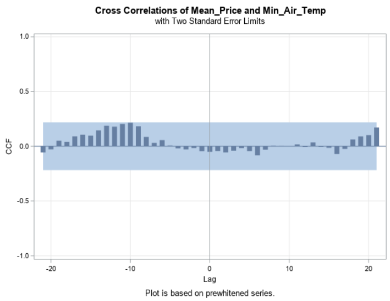
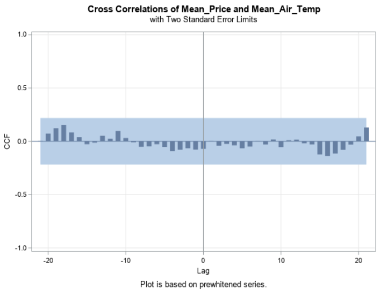
1. Mean\_Precipitation

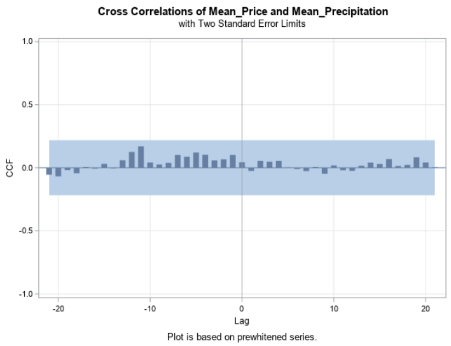
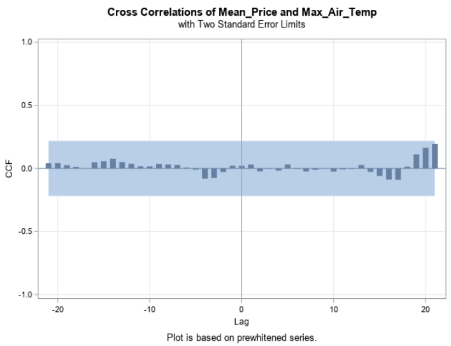
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| AR | MA | AR(S) | MA(S) | AIC | SBC | WN |
| 1 | 1 | 0 | 0 | -97.7031 | -90.4107 | No |
| 2 | 0 | 0 | 0 | -98.4626 | -91.1701 | No |
| 2 | 1 | 0 | 0 | -104.747 | -95.0242 | Yes |
| 0 | 2 | 0 | 0 | -98.225 | -90.9326 | No |
| 1 | 2 | 0 | 0 | -96.2936 | -86.5703 | No |
| 2 | 2 | 0 | 0 | -94.2979 | -82.1438 | No |

For Mean\_Precipitation, the model with AR(2), MA(1), seasonal AR(0), and seasonal MA(0) components was selected.

**Cross Correlation Analysis after prewhitening:**

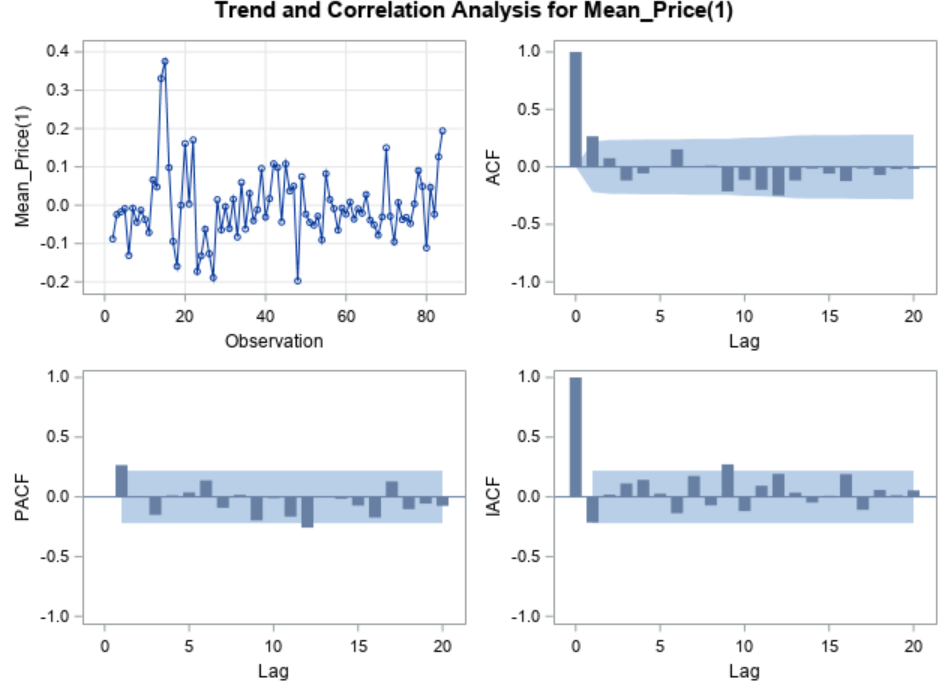
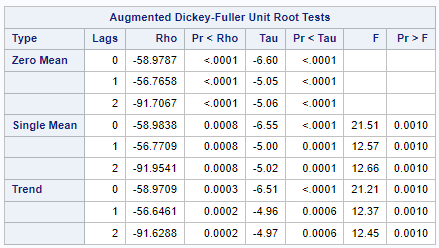
No lag effect was observed even after conducting prewhitening.





**Model Selection:**

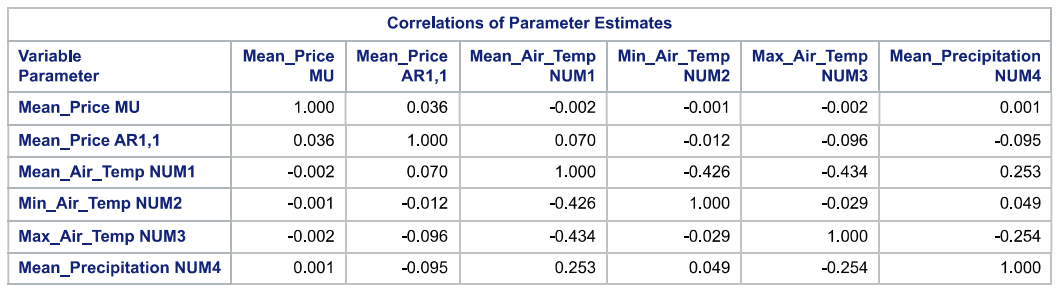
After exploring our variables and conducting cross-correlation analysis, our team decided to test multiple models and selected the one with the best accuracy. However, before delving into the model, we revisited the target variable's ACF and PACF. According to the ACF of the Mean\_value with one-order differentiation, we observed a pattern of gradually declining autocorrelation as lags increase. On the other hand, the PACF plot showed that the first lag is the only clearly significant lag, implying an autoregressive order of 1. Therefore, we initiated the process by applying an AR (p=1) model. Furthermore, a unit root test for the Mean\_Price indicated that the target variable, after one-order differentiation, is stationary.

Our first model was ARIMA(1,1,0) and contained all X variables (Max\_Air\_Temp, Min\_Air\_Temp, Mean\_Air\_Temp and Mean\_Precipitation). We chose a holdback period of 12 and this model produced the following summary statistics: MAPE: 8.62%, MAE 0.096133, MSE: .02221133, RMSE: 0.14877, AIC: -149.048, SBC: -134.535 and had White Noise Residuals.

Our second model that we tested was ARIMA(2,1,0) and contained all X variables (Max\_Air\_Temp, Min\_Air\_Temp, Mean\_Air\_Temp and Mean\_Precipitation). We chose a holdback period of 12 and this model produced the following summary statistics: MAPE: 8.05%, MAE 0.089978, MSE: 0.020043, RMSE: 0.14157, AIC: -147.159, SBC: -130.228 and had White Noise Residuals.

Up until this point, we computed statistics for AR1 and AR2 models to make comparisons, as the data remained the same with no changes in the x variables. During this process, we recognized that using multiple temperature variables might lead to multicollinearity among them, marked by moderate correlation, unlike precipitation.



As a result, we decided to experiment with models containing fewer x variables to create a less complex model. We tested models with a single x variable, and the best-performing ones were Max\_Air\_Temp and Mean\_Precipitation, respectively.

The third model we tested was ARIMA(1,1,0) with only Max\_Air\_Temp as an input variable. We chose a holdback period of 12 and this model produced the following summary statistics: MAPE: 7.54%, MAE 0.084292, MSE: 0.018702, RMSE: 0.13675, AIC: -154.987, SBC: -147.731 and had White Noise Residuals.

The fourth model we tested was ARIMA(1,1,0) with only Mean\_Precipitation as the input variable. We chose a holdback period of 12 and this model produced the following summary statistics: MAPE: 7.88%, MAE 0.08751, MSE: 0.0188341, RMSE: 0.13543, AIC: -153.504, SBC: -146.248 and had White Noise Residuals.

We also explored various combinations of temperature and precipitation, such as "Min\_Air\_Temp and Mean\_Precipitation" and "Mean\_Air\_Temp and Mean\_Precipitation". Among these combinations, the most effective one was found to be Mean\_Air\_Temp and Mean\_Precipitation. Ultimately, the accuracy of our model became a concern, as it exhibited signs of underfitting. To address this issue, we examined whether the predicted values lagged behind the actual values. We ruled out this possibility by analyzing an out-of-sample forecast graph.

The final model we tested was ARIMA(1,1,0) with Mean\_Air\_Temp and Mean\_Precipitation as input variables. We chose a holdback period of 12 and this model produced the following summary statistics: MAPE: 7.74%, MAE 0.086097, MSE:0.018104, RMSE: 0.13455, AIC: -151.694, SBC: -142.019 and had White Noise Residuals.

Below is a table of summary statistics for each model:

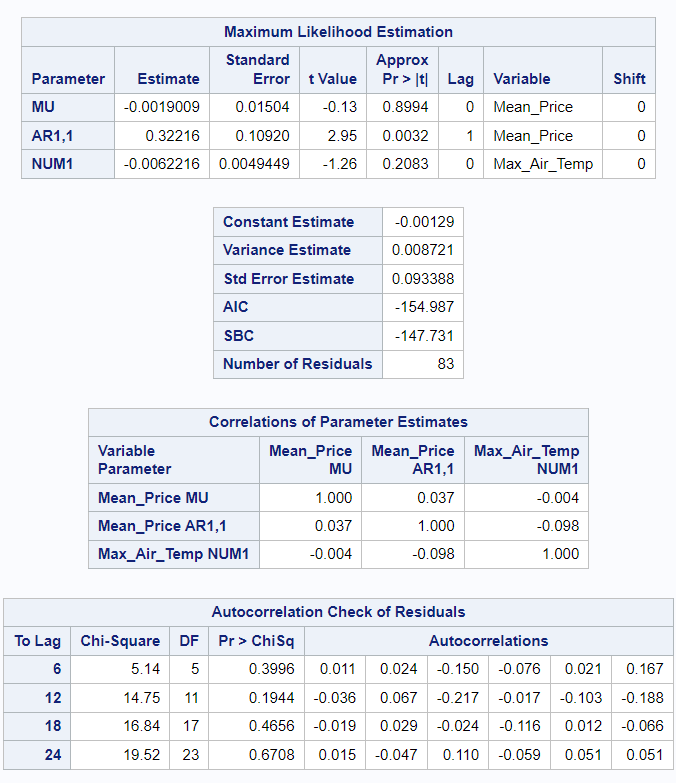
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | AR (p) | I (d) | MA (q) | Holdback Period | MAPE | MAE | MSE | RMSE | AIC | SBC | Residual  White Noise |
| ALL X | 1 | 1 | 0 | 12 | 8.62% | 0.096133 | 0.022133 | 0.14877 | -149.048 | -134.535 | Yes |
| ALL X | 2 | 1 | 0 | 12 | 8.05% | 0.089978 | 0.020043 | 0.14157 | -147.159 | -130.228 | Yes |
| MAX\_TEMP | 1 | 1 | 0 | 12 | 7.54% | 0.084292 | 0.018702 | 0.13675 | -154.987 | -147.731 | Yes |
| MEAN\_PRECIP | 1 | 1 | 0 | 12 | 7.88% | 0.08751 | 0.018341 | 0.13543 | -153.504 | -146.248 | Yes |
| MEAN\_TEMP & Mean\_Precipiation | 1 | 1 | 0 | 12 | 7.74% | 0.086097 | 0.018104 | 0.13455 | -151.694 | -142.019 | Yes |

**Best Model:**

After creating multiple models, we chose our third model, ARIMA(1,1,0) with Max\_Air\_Temp as X variable for our final model. One of the primary factors in the best model selection was model accuracy. To gauge this, we evaluated several key metrics including Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). For each of these statistical measures, we wanted the model with the lowest value. Even though the models produced similar results, our third model had some of the lowest values with the least amount of variables.

Next, we looked at model simplicity, while model accuracy is important, it's also essential to consider model simplicity. A simpler model with fewer variables is often preferred because it is more interpretable and less prone to overfitting. The third model, with Max\_Air\_Temp as the only exogenous variable, was relatively straightforward and less complex. Furthermore, a simpler model is advantageous because it requires fewer data inputs and is easier to explain to stakeholders who may not have a deep understanding of advanced modeling techniques.

We then looked at the model fit statistics of Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC). Again, the models all produce similar results but we confirmed our selection of the third model since it had the lowest or close to the lowest SBC and AIC values. Finally, to confirm our model selection once more, we checked that the residuals from our model exhibited white noise. All of our models exhibited this behavior which shows that they all captured the underlying patterns in the data well.

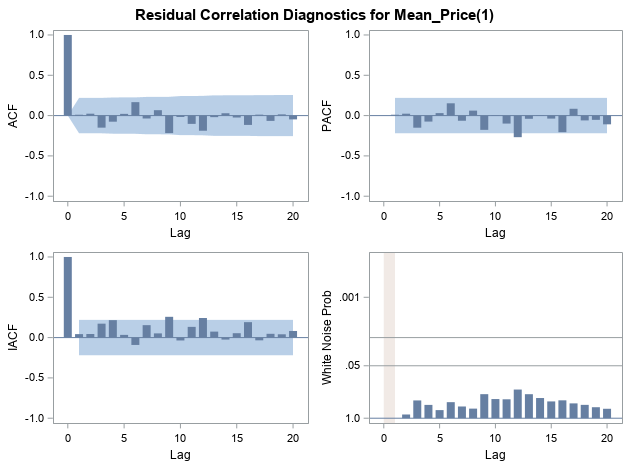


**Maximum Likelihood Estimation:**

"MU" represents the intercept, and its estimate is approximately -0.0019009. "AR1,1" is the autoregressive term with an estimate of 0.32216, indicating a positive autocorrelation with a lag of 1. "NUM1" represents the coefficient for Max\_Air\_Temp with an estimate of -0.0062216, suggesting that a one-unit increase in Max\_Air\_Temp is associated with a decrease of approximately 0.6 cents in the global coffee price..

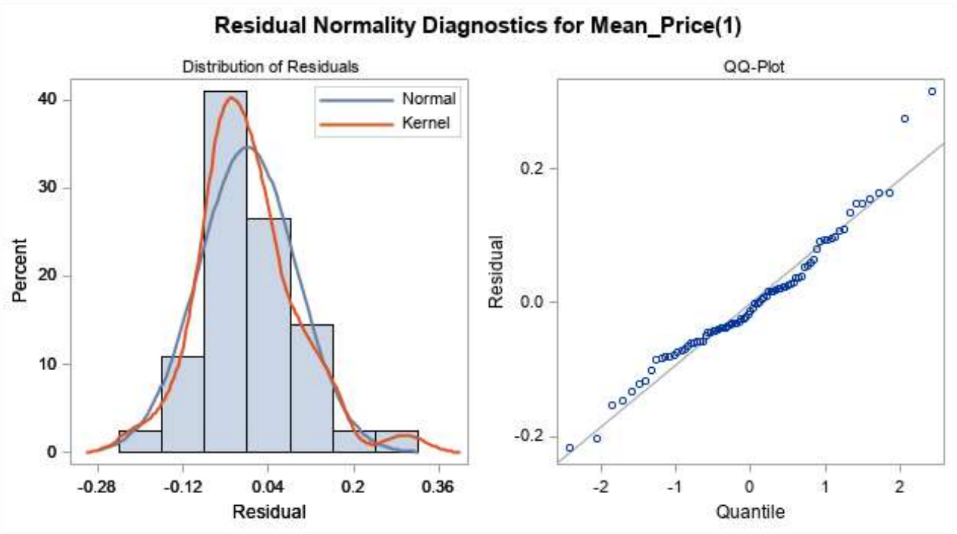
**Autocorrelation Check of Residuals:**

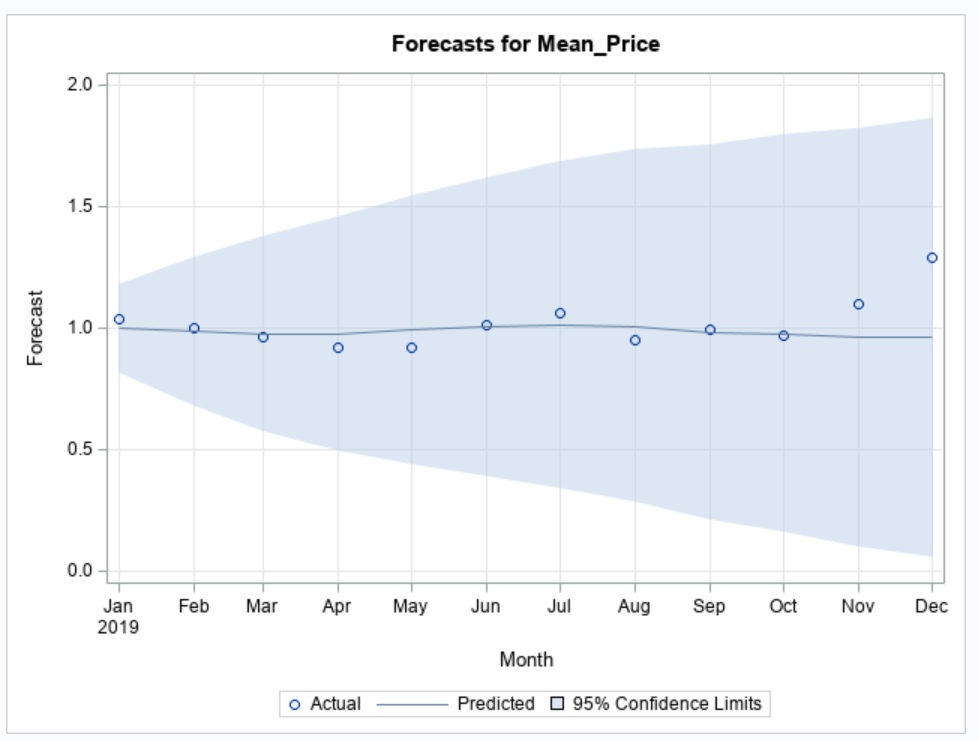
For each lag, the Chi-Square statistic tests whether the autocorrelations are significantly different from zero. High p-values (close to 1) suggest no significant autocorrelation. The autocorrelations for the various lags are presented. None of them show strong correlations, as indicated by the relatively low absolute values of the autocorrelations. This suggests that the residuals do not exhibit strong autocorrelation.



**Residual Correlation Diagnostics for Mean\_Price(1):**

There are no significant lags in the ACF, PACF, and IACF, and we observe white noise in the data. Hence, we can conclude the model does not show significant autocorrelation patterns, and the data behaves like random noise with no discernible structure or trends.

The residuals are relatively normally distributed.



Our best model has demonstrated its forecasting capabilities. We set a holdout period of 12 months, and it appears to accurately forecast the holdout set. The forecasted points all fall within the confidence interval, with the majority closely aligning with the actual values.

**Conclusion:**

While the final model demonstrates reasonable forecasting accuracy, it does not establish a significant correlation between climate features and coffee prices. Here are possible reasons for this result:

* Threshold Effects: Extreme weather events like severe droughts or frost have a greater likelihood of significantly impacting coffee prices, while minor weather fluctuations may have negligible effects.
* Resilience to Moderate Variations: Coffee plants in the MG region may be relatively resilient to moderate variations in climate features like precipitation and air temperatures. Some regions, like MG, might have coffee varieties or cultivation practices adapted to local climate conditions, making them less sensitive to minor fluctuations.
* Market Factors: Other factors, such as global supply and demand dynamics, geopolitical events, currency fluctuations, and consumer preferences, can play a substantial role in coffee price movements. These market factors may overshadow the influence of climate variables in short-term price fluctuations.

**Recommendations:**

After completing this forecasting project, our team has reflected on what could be done in the future to improve this report. Our first recommendation is to limit the limitations that presented itself in this assignment. We faced difficulties from our massive dataset with an abundance of years, states, regions, stations, etc. We spent a lot of time cleaning and preparing the data that in turn limited our time to build an accurate model. If we were to restart this assignment, we would have chosen more specific data that had more valuable input variables. Another recommendation is to not forward-fill the values for missing data. Since we transformed both datasets from either daily to monthly or from hourly to monthly, we could have just used an average or minimum or maximum value for that month. This would help eliminate extra noise that this process may have caused.

If time was not a factor, increasing the span of the dataset may have provided a more accurate model. Since the objective was to see how climate change had affected the price of coffee, six years is not enough time to show the change in temperature caused by climate change. Increasing the timespan to include the 1970’s or 80’s would show a slight trend in the temperature from an annual average temperature of approximately 27 degrees celsius for Central/South America to 27.6 degrees celsius in 2022. An increase of 0.6 degrees celsius may not seem like a lot but could greatly impact the growing seasons of different crops such as coffee. Based on the model we created from our current dataset, the more impactful trend would be the change in rainfall from the 1980’s to 2022.

We could have delved deeper into the project by conducting a comprehensive examination of how climate change affects coffee production. This could involve utilizing climate models and historical climate data to quantitatively evaluate the correlation between temperature and precipitation variations and their impact on coffee yield and quality.

Citations

Adam. “Brazilian Coffee Facts.” *Green Farm Coffee Company*, 9 Jan. 2017, www.greenfarmcoffee.co.uk/brazil-coffee-facts/#:~:text=80%25%20of%20Brazil’s%20coffee%20is,is%20between%20May%20and%20September.&text=There%20are%20roughly%20220%2C000%20coffee,for%2063.9%25%20of%20total%20exports.

“Brazil Coffee Route of Minas Gerais: Tips to Plan Your Trip.” *Food’n Road*, 8 Nov. 2021, foodandroad.com/brazilian-coffee-route-of-minas-gerais/.

“Brazil Coffee.” *Espresso International*, 23 Jan. 2021, www.espresso-international.com/brazil-coffee#:~:text=Today%2C%20coffee%20is%20only%20responsible,rich%20and%20important%20coffee%20culture.

“Brazil.” *Melbourne Coffee Merchants*, 6 June 2023, melbournecoffeemerchants.com.au/origin/brazil/#:~:text=Brazil%20is%20the%20world’s%20largest,at%20least%2070%25%20is%20Arabica.

Charles, Sarah. “How More Accurate Weather Forecasting Could Stabilize Coffee Prices.” *Coffee Intelligence*, 19 July 2023, intelligence.coffee/2023/07/weather-forecasting-coffee-prices/#:~:text=Poor%20weather%20conditions%20can%20significantly,the%20cost%20of%20green%20coffee.

Costa, Bruna. “A Concise Guide to Brazil’s Major Coffee-Producing Regions.” *Perfect Daily Grind*, Perfect Daily Grind, 25 May 2021, perfectdailygrind.com/2016/04/a-concise-guide-to-brazils-major-coffee-producing-regions/.

“Global Warming by Continents.” *Worlddata.Info*, www.worlddata.info/global-warming.php. Accessed 10 Oct. 2023.

Koons, Matt. “How Does Altitude Affect Coffee Roasting and Taste?” *Fathom Coffee Roasters - A Deeper Love For Coffee*, 9 Aug. 2023, fathomcoffee.com/high-altitude-coffee/#:~:text=Farmers%20typically%20grow%20Arabica%20coffee,and%20the%20Tropic%20of%20Capricorn.