BUSINESS REPORT

Business Forecasting Exam

Aarhus University
Flow ID: 95

Executive Summary

After conducting an in-depth analysis of the sales performance data for Brew Station's coffee vending machine in Vinnytsia, Ukraine, it has been observed that weather conditions, particularly average temperature, have a significant impact on coffee sales. This finding was supported by running a statistical regression model, which indicated a clear correlation between average temperature and sales volume.

The analysis supports the rationale of the entrepreneur that the colder days tend to generate higher demand for hot beverages like coffee, while warmer days may lead to a reduction in sales due to decreased interest in hot drinks.

Based on the findings and the influence of average temperature on coffee sales, it is recommended to further explore potential relocation of the vending machine to a site with more favorable weather conditions to maximize sales potential and drive business growth.

Table of Contents

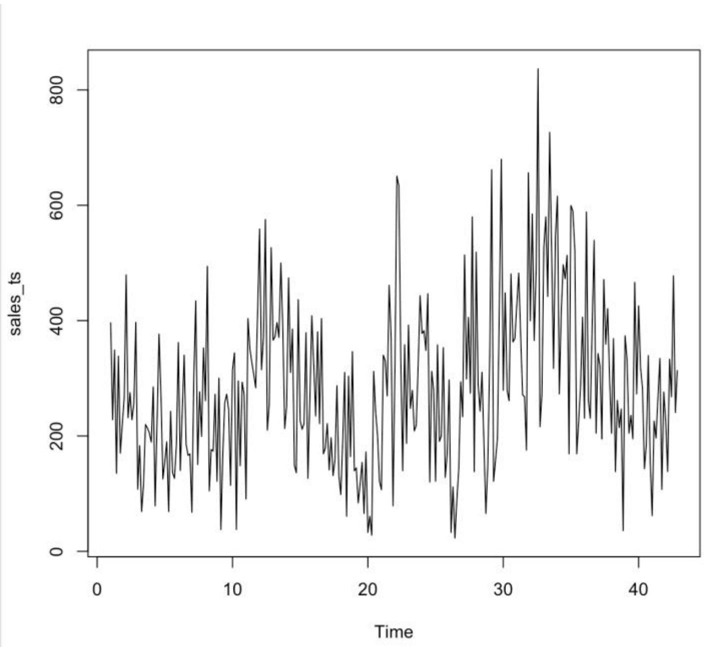
2. Analysis and Findings

For this analysis, we are dealing with daily data from a Coffee shop. This means that the frequency will be set to 7 as we are expecting weekdays to be similar to each other in terms of sales (which is the main variable of interest). We have 294 data point available for the analysis starting on March 2024 and ends on the 23rd of December 2024.

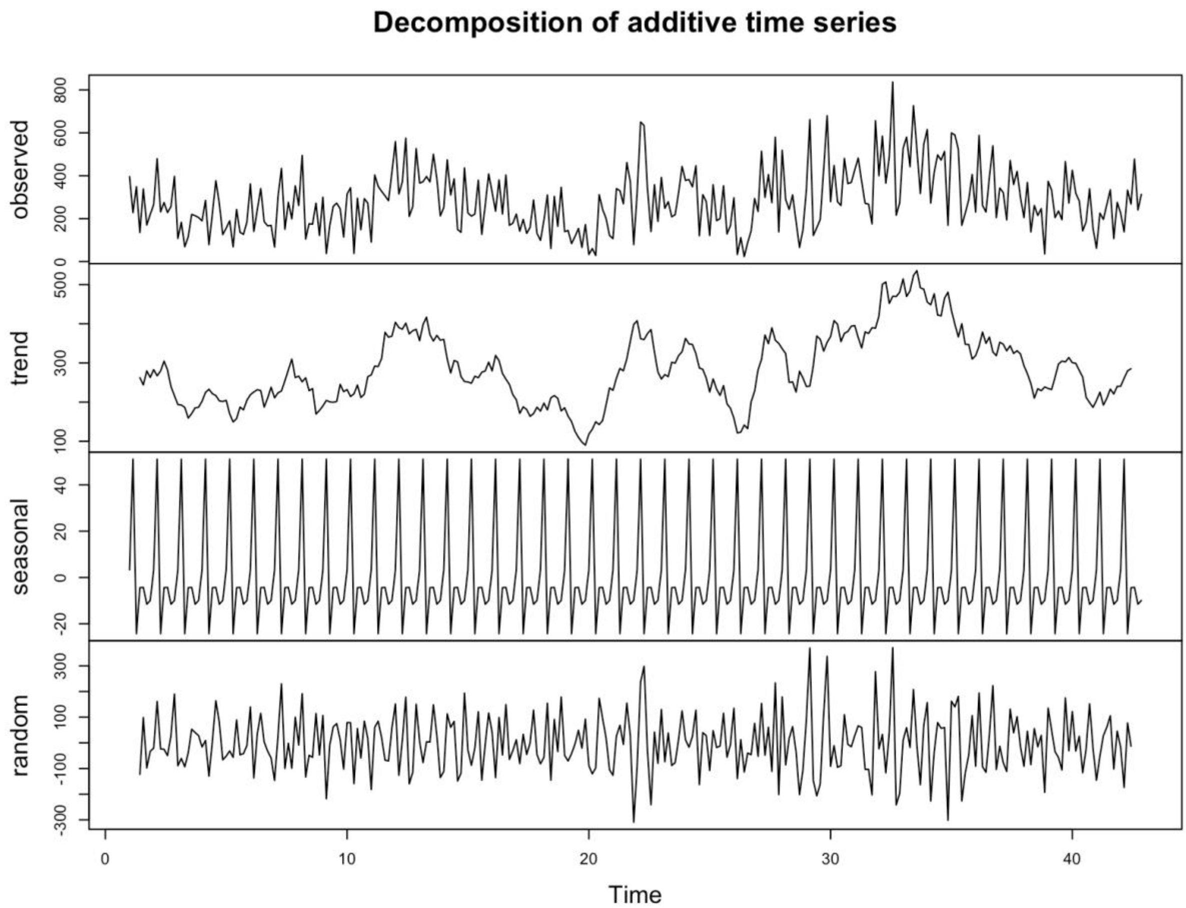
2.1 Decomposition of Daily Sales

Based on the entrepreneur's interest to find out what factors impact the sales throughout the year, and the task to forecast sales for the remainder of the year, the Total Daily Sales Revenue has been identified as the key variable of interest for this analysis.

To examine the seasonality and trend in the data, the time-series decomposition will be used.



Based on visual inspection of the time series plot for the sales, the data exhibits relatively stable variance over time. This characteristic suggests an additive decomposition would be most appropriate, as the seasonal fluctuations appear constant over time.



The time series decomposition plot reveals that total daily revenue does exhibit slight upward trend with quite high volatility. The sales seems to be increasing over time throughout the year peaking at the end of the year, but the trend is not strictly upward and tends to be volatile throughout the year.

Looking at the seasonality component in the plot we can see a clear seasonal patterns with peaks occurring on a weekly basis. This indicates weekly seasonality is present in the data. The larger peaks appearing throughout the year (as visible on the trend component) could represent long-term seasonality. For instance the sales will be most likely peaking in the winter as the temperature goes down. In order to assess the long term seasonality, we would need data for several years to make stronger conclusions.

Now, looking into the decomposition using regression:

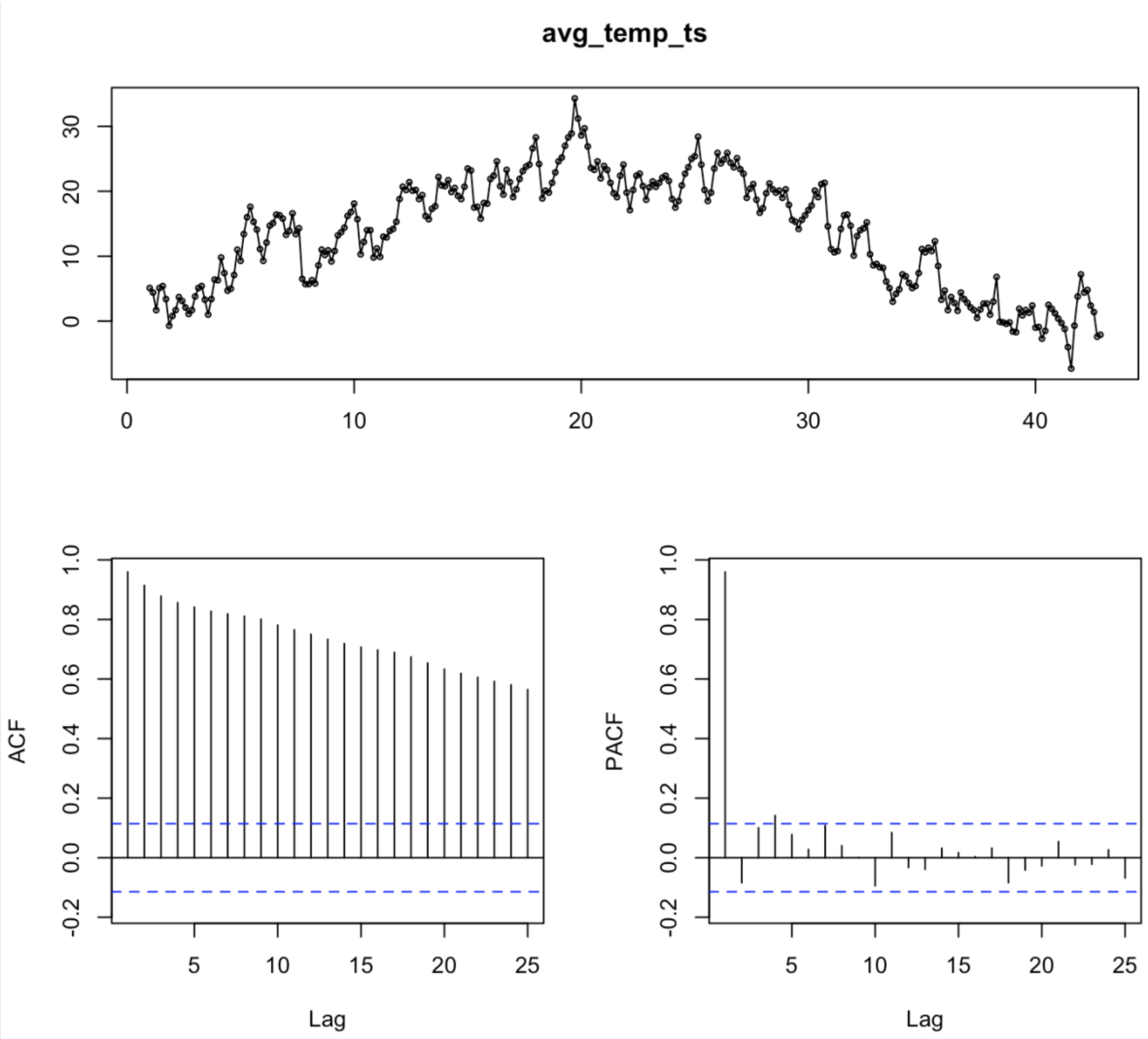
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              215.55932   26.33040   8.187 8.97e-15 ***
month_dummies_seasonalS1  17.65773   30.93459   0.571   0.5686    
month_dummies_seasonalS2  60.41533   30.93290   1.953   0.0518 .  
month_dummies_seasonalS3 -10.49040   30.93152  -0.339   0.7347    
month_dummies_seasonalS4   6.69196   30.93045   0.216   0.8289    
month_dummies_seasonalS5  10.56194   30.92968   0.341   0.7330    
month_dummies_seasonalS6  -2.90570   30.92922  -0.094   0.9252    
time_trend_seasonal        0.38811    0.09742   3.984 8.61e-05 ***
```

The trend coefficient is positive at 0.388 and very significant p-value indicating a significant long-term increase in the series. This suggests that, on average, the sales increases by 0.388 units per time period, holding seasonal effects constant.

Looking at the seasonal dummy coefficients, there aren't any significant one's present. The dummy number 2 is statistically closest to the significance with p-value of 0.0518, but overall the weekly seasonality does not seem to be present in the data.

2.2 Dynamic properties of Average Temperature

Considering entrepreneur's hypothesis that average temperature has an impact on the daily sales, I will select this variable as the primary focus for this analysis. The prior knowledge of the entrepreneur suggests that colder days drive higher demand for hot drinks, while warmer days are reducing it.



Looking at the plot above, the average temperature shows clear trend in the data. Firstly as the year goes into the summer the temperature is increasing and then starts to decrease in the winter. This behavior is expected and leads to time dependencies in the data. This is visible looking at the ACF plot that decays slowly with time – this suggests that the series is not-stationary.

```
        Augmented Dickey-Fuller Test

data:  avg_temp_ts
Dickey-Fuller = -1.896, Lag order = 6, p-value = 0.6201
alternative hypothesis: stationary
```

The non-satationarity is confirmed statistically with Augmented Dickey-Fuller test where we were not able to reject null hypothesis of unit root.

For further analysis some of the model like regression will require the average temperature to be stationary, so we will need to assess whether to apply differencing to this variable.

2.3 Estimating Relationships with Regression

Before running regression on Sales and Average temperature, we need to make sure the autocorrelation is not present in the variables. The ADF test earlier showed that average temperature is non-stationary.

```
        Augmented Dickey-Fuller Test

data:  sales_ts
Dickey-Fuller = -3.5675, Lag order = 6, p-value = 0.03662
alternative hypothesis: stationary
```

Running ADF test on Sales (figure above), we were able to reject the null-hypothesis suggesting that the series is stationary.

```
        Phillips-Ouliaris Cointegration Test

data:  combined_series
Phillips-Ouliaris demeaned = -220.39, Truncation lag parameter = 2, p-value = 0.01
```

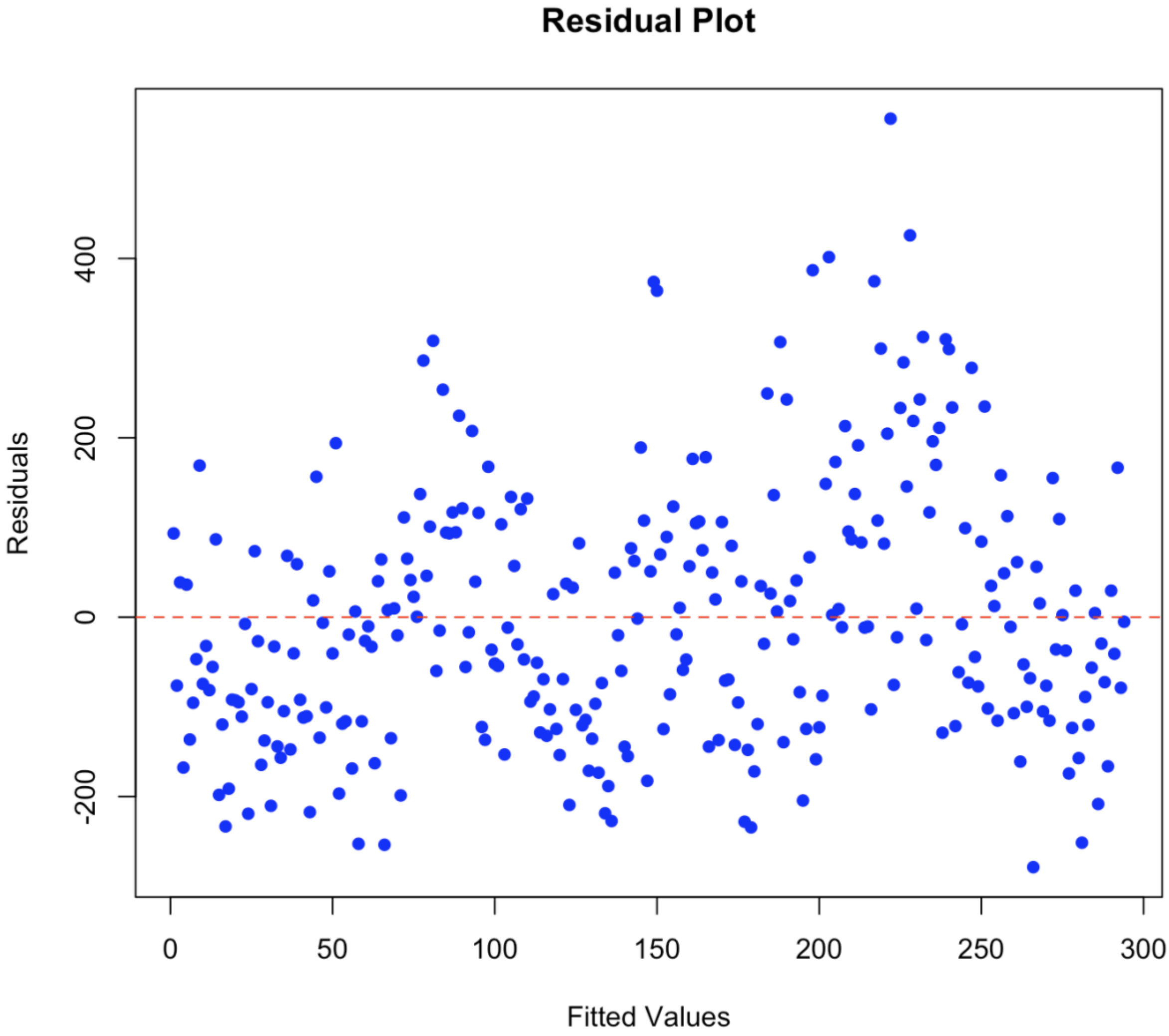
Given that one of the series is non-stationary, we test for cointegration using the Phillips-Ouliaris test. The test rejects the null hypothesis of no cointegration at the 0.05% significance level, suggesting that we can proceed with level variables for this regression.

Given the setup of the information available, I decide to run static regression as I expect the weather today will be the best proxy for sales today.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 314.0527    15.6600  20.055   <2e-16 ***
avg_temp_ts  -2.1869     0.9772  -2.238    0.026 *
```

The summary of the simple regression shows the intercept is highly significant as well as the average temperature.

The beta coefficient for average temperature suggests that for one unit increase in temperature, the sales go down by -2.1869 showing a negative relationship between these two variables. This confirms the hypothesis of the entrepreneur that sales go down with increasing temperature.



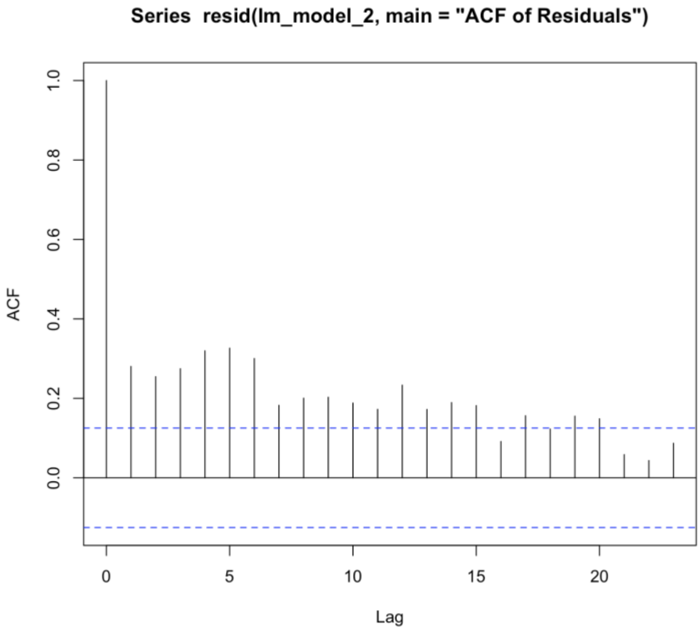

```
        Augmented Dickey-Fuller Test

data:  resid(lm_model)
Dickey-Fuller = -3.438, Lag order = 6, p-value = 0.04906
alternative hypothesis: stationary
```

The residuals does not seem to exhibit any strong patterns and the ADF test on residuals rejects H0, suggesting there is no autocorrelation in the residuals. This means that the temperature captures the movements of sales well.

2.4 Regression performance on Out of Sample data

After splitting the data into insample and out of sample. The checks for cointegration have been made again, which showed that the insample variables are still co-integrated, so we proceeded with running regression model on levels of out of sample data. The newly obtained coefficients align with the previous observations we made. The average temperature still acts as a significant variable (figure below).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   350.228     23.161  15.122   <2e-16 ***
insample_temp  -4.058      1.323  -3.067   0.0024 **
---
```



Looking at the ACF plot of residuals, there is multiple significant spikes that decay towards zero quite slowly, which shows there could be a violation of an assumption of uncorrelated residuals.

```
        Augmented Dickey-Fuller Test

data:  resid(lm_model_2)
Dickey-Fuller = -3.7317, Lag order = 6, p-value = 0.023
alternative hypothesis: stationary
```

However ADF test rejected the H0, which means that statistically the residuals are stationary.

```
> reg_acc
                ME     RMSE    MAE       MPE     MAPE        ACF1 Theil's U
Test set -65.39809 132.8966 110.45 -62.66771 71.98938 0.006058751 0.4103642
```

In terms, of out of sample performance, the RMSE is 132.89.

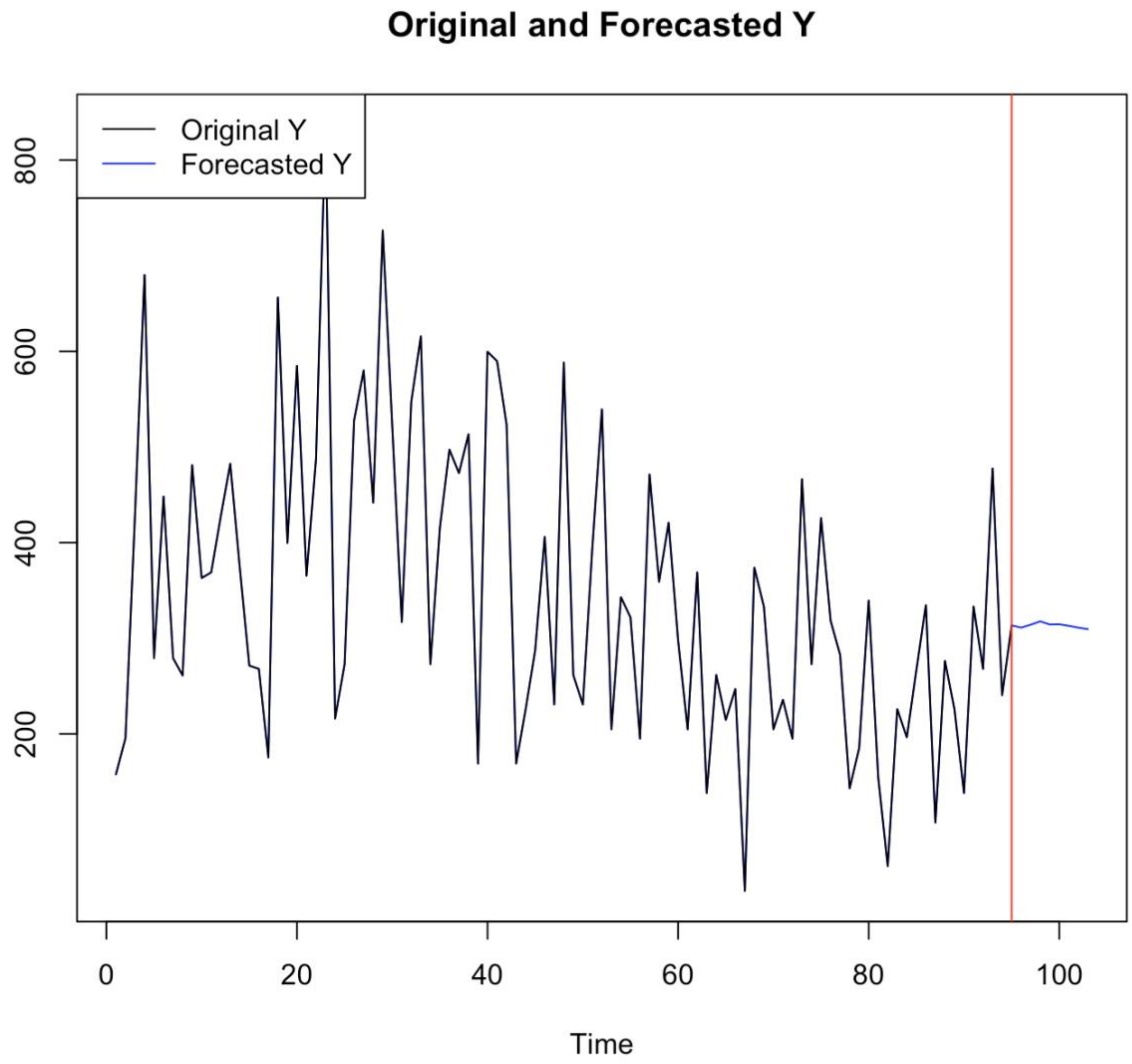
Both regression lead to a conclusion that average daily temperature has an significant impact on daily sales

2.5 Forecast for December using regression

The end of the year forecast using the regression model fitting Sales and Average temperature yield following values:

```
       1        2        3        4        5        6        7        8
311.1369 314.1803 317.6338 314.4172 314.5995 313.0140 311.1005 309.5696
```

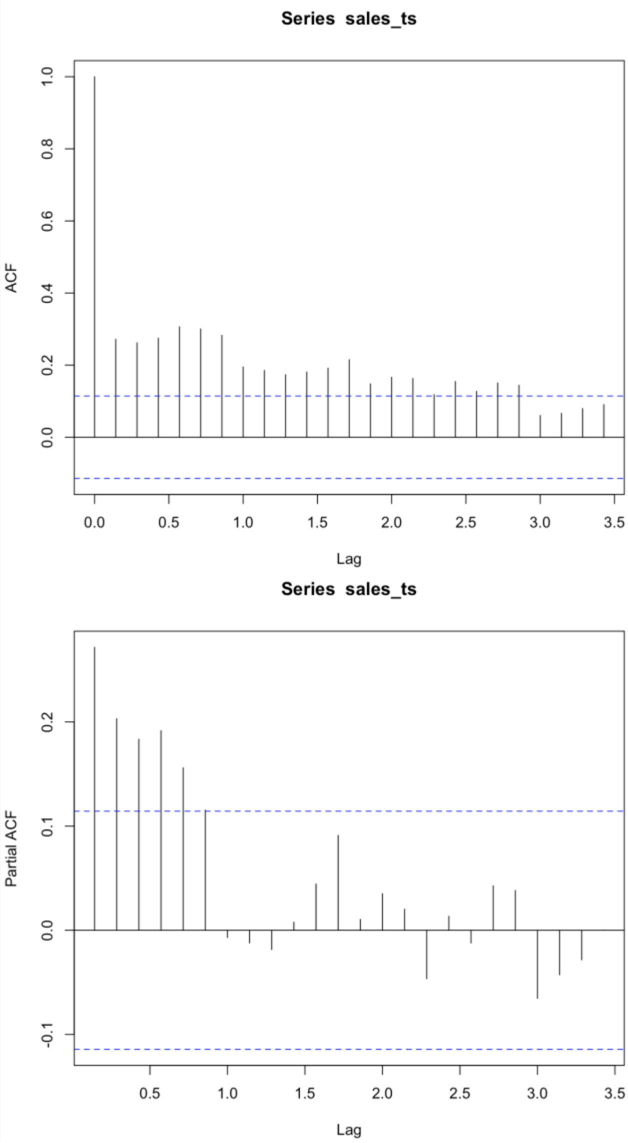
Here are the forecasted values visualized in the plot:



2.6 Arima model

I decided to run auto-arima, which suggested ARIMA(1,1,1).

```
Series: insample_sales
ARIMA(1,1,1)

Coefficients:
          ar1      ma1
      -0.0073  -0.8411
s.e.   0.0780   0.0454

sigma^2 = 18344:  log likelihood = -1543.52
AIC=3093.04   AICc=3093.14   BIC=3103.53

Training set error measures:
                     ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 1.836873 134.6097 106.8879 -34.74844 60.23715 0.7038759 -0.001081493
```



Looking at the ACF and PACF, both have significants spikes which decline towards zero slowly, but ADF test on Sales showed that the series is stationary, which would mean an ARMA(1,1) model being appropriate, so the auto arima more less aligns with the patterns.

ARIMA(1,1,1) test set accuracy:

```
                ME     RMSE      MAE       MPE    MAPE        ACF1 Theil's U
Test set -105.7624 156.4094 134.2236 -81.45166 87.0595 -0.02962423  0.489979
```

The RMSE of this arima is 156.41, which shows a worse out of sample performance compared to the regression model.

2.7 Forecast of Arima for rest of December

ARIMA(1,1,1) December forecast:

```
> manual_arima_forecast_mean
[1] 385.7028 385.8509 385.8498 385.8498 385.8498 385.8498 385.8498 385.8498
```

Regression December forecast:

```
> lm_model_forecast_mean
       1        2        3        4        5        6        7        8 
311.1369 314.1803 317.6338 314.4172 314.5995 313.0140 311.1005 309.5696
```

Comparing the forecast, we can see that Regression underestimates the amount of sales compared to ARIMA. Furthermore, ARIMA generated very linear estimations with low variance suggesting that Regression might be a better fit for our data.

2.8 Comparing Regression and Arima

```
	Diebold-Mariano Test

data:  residuals(auto_arima_forecast)residuals(forecast_lm_model_2)
DM = -1.7083, Forecast horizon = 49, Loss function power = 2, p-value = 0.08886
alternative hypothesis: two.sided

> dm_test_errors

	Diebold-Mariano Test

data:  auto_arima_forecast$mean - outsample_salesforecast_lm_model_2$mean - outsample_sales
DM = 0, Forecast horizon = 49, Loss function power = 2, p-value = 1
alternative hypothesis: two.sided
```

Comparing both models using Diebold-Mariano test for residuals and errors, we cannot reject the null hypothesis and therefore conclude that the forecasts are of equal predictive ability. This means that the models are ideal to be combined

2.9 Forecast Combinations

Looking at the RMSE of all models, including Equal Weights and Granger-Ramathan combination, we see that the lowest RMSE was obtained combining the models with Granger-Ramathan method:

```
> rmse_results_df
  regression_outsample_rmse arima_outsample_rmse equal_weights_rmse  gr_rmse
1                  132.8966             156.4094           143.6194 114.9871
>
```

2.10 Final Forecast

Producing the forecast for the end of the year, I chose the combined model using Granger-Ramathan method as it showed the lowest out of sample RMSE. The weitghts of the model are displayed in the model summary below:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.416e+05  3.206e+05  -0.442    0.661
auto_arima_forecast$mean  3.678e+02  8.311e+02   0.443    0.660
forecast_lm_model_2$mean -9.613e-02  1.608e+00  -0.060    0.953
```

I ran both ARIMA and Regression predictions for last 8 days of the year and then combined both using the weights obtained from the coefficients of Granger-Ramathan model. Using this formula:

Combined_forecast = (-1.416e+05) + (3.678e+02 * arima_forecast) + (-9.613e-02 * regression_forecast)

This yields following values for the end of the year:

```
> Combined_forecast
       1        2        3        4        5        6        7        8 
231.5922 285.7572 285.0295 285.3416 285.3241 285.4765 285.6604 285.8076 
```

Recommendation

Given the substantial influence of average temperature on coffee sales, one key recommendation is to consider relocating the vending machine to a site characterized by more favorable weather conditions. By identifying a location that aligns with the sales-driving factor of colder temperatures, Brew Station can potentially maximize sales potential and stimulate business growth.

Conclusion

In conclusion, the analysis underscores the importance of weather conditions, particularly average temperature, in shaping coffee sales trends for Brew Station's vending machine in Vinnytsia, Ukraine. Leveraging insights derived from the regression model and the Granger-Ramanathan method, the forecasted sales figures provide valuable guidance for strategic decision-making. Moving forward, by exploring the recommendation to explore relocation to a more fitting site, Brew Station can achieve more favorable business outcomes.