Executive Summary for Time Series and Parametric Analysis on Sales

Gooden, Nina S. | Student ID #: 009823504

Western Governors University

Dr. William Sewell

December 23, 2022

Table of Contents

[Statement of the Problem and the Hypothesis 3](#_Toc123520615)

[Summary of the Data-analysis Process 4](#_Toc123520616)

[Outline of the Findings 6](#_Toc123520617)

[Technique and Tool Limitations 7](#_Toc123520618)

[Summary of Proposed Actions 7](#_Toc123520619)

[Benefits of the Study 8](#_Toc123520620)

[References 9](#_Toc123520621)

# Statement of the Problem and the Hypothesis

During the recent pandemic, Montgomery County of Maryland saw record fluctuations in their warehouse and retail sales. While entire industries shrank and contracted, Montgomery County stores saw sales growth ranging as high as a 288% increase (Longo & Fischer, 2020). However, the true impact of those sales—and the impact of sales after 2020—remains uncertain.

This study will utilize autoregressive integrated moving average (ARIMA) to describe autocorrelations in the data and create a time series model which will predict product sales within the county during the height of Montgomery County’s pandemic sales. In doing so, the model will clarify the deviation between what the model predicted through historical data, as opposed to what occurred in the post-pandemic market. Additionally, by creating this model, the county will be better equipped to measure the overall impact of market downturn, plan for future market adjustments, and compare linear predictions to actual performance.

The objective of building a time series model is to “extrapolate the dynamic pattern in the data for forecasting future observations, to estimate the effect of known exogenous interventions, and to detect unsuspected interventions” (Tiao, 2015) according to the needs of the entity the model is created for. In this case, that need is to use historical data to allow county leaders to make informed decisions on continued supply chain needs, purchasing, and marketing efforts in the post-pandemic market.

To demonstrate the merit of the time series analysis, the created model will be trained and tested with holdover data in order to reach a 70% or greater accuracy, derived from the statistical analysis t-test. The study will be deemed successful, and the alternative hypothesis H1 accepted if the if the time series model is created with a 70% confidence score. If 70% cannot be reached, H0 will not be rejected and additional work will be necessary to evaluate the stated problem.

# Summary of the Data-analysis Process

The primary goal of this analysis is to identify patterns in sales data to build a forecast model for the providing entity. The data was last updated December 9, 2022 and is updated monthly. There are 307,646 records in the data set and 9 columns. Overall data sparsity is < 5%.

The data-analysis process was structured around the industry-accepted steps to ensure analysis is accurate and reproducible.

First, efforts were made to ensure the tools and settings utilized fit with data preparation standards. To align with these efforts, the analysis began in Jupyter Notebook, via the text editing capabilities of Visual Studio Code. Packages and libraries were imported into the environment as follows:

* pandas as pd, used for DataFrame components and data manipulation
* numpy as np, utilized for array work and base Python interactions
* matplotlib.pyplot as plt, and seaborn as sns used for data visualizations
* sklearn used for complex mathematical components, as well as testing model accuracy
* statsmodels.api as sm, used as the primary interface for the ARIMA modeling and data preparation
* datetime, use for data preparation

After importing the dataset, exploratory functions were launched in order to investigate the data further. Columns unnecessary for the evaluation were dropped.

The fillna and append methods were used to input missing values in the ‘item\_type’ and ‘date’ columns respectively. Necessary date-time columns were created in preparation for ARIMA functionality. Finally, the date data was set at the index for the DataFrame and entries that shared a date were combined with the sum and round functions and organized in ascending order.

After all the duplicates were removed, there were only 43 entries left in the DataFrame. This meant that the model would be more sensitive to each fluctuation and would result in large variances.

Augmented Dickey-Fuller unit root tests were deployed in order to check for stationarity. The data was transformed through differencing and reevaluated.

The results of the second test resulted in the rejection of the null hypothesis, as the absolute value of the ADF Statistic value was greater than the critical value. In addition, the p-Value was below 0.05, confirming that the data was stationary.

In order to run the autocorrelation analysis and the partial autocorrelation analysis, researchers used the train\_test\_split function on the differenced data to split the data into two groups, using the 4:1 ratio.

The acf function was applied to the train dataset in order to correlate the points of the dataset with a lagged version of itself. The autocorrelation function begins at lag 0 and results in a correlation of 1. The pacf function was then applied to the train dataset in order to correlate the points of the dataset by each successive lagged term.

Pre-modeling assessment was conducted with visualizations of seasonal decomposition and spectral density analysis. Both were found to exhibit no evidence of seasonal trends.

The auto\_arima function was used to test different arguments for the potential ARIMA model, then verified manually by performing a stepwise search for the minimum Akaike's Information Criterion (AIC) possible. The returned best model values were ARIMA(2,0,0)(0,0,0)[0] intercept. When run, the SARIMAX Results were within acceptable ranges, with p-values that were not zero, but largely less than .05.

With the model created, the data was split again, and the updated test and train models were filtered into the predict function in order to produce prediction information on the train and test values.

# Outline of the Findings

The research question was whether or not a time series model could be used to predict the next quarter’s sales data for each item type in the dataset. While there was not enough support to itemize the prediction, a time series model was created for overall sales with 70% confidence interval. This model was able to create a predictive model for the combined warehouse and retail sales for an additional quarter.

The model predicted that overall trend in the data was steadily increasing. However, there was a considerable downturn approaching. The implication of this downward trend was that the county should have suffered from sales losses at the beginning of the year.

However, when looking at the existing data, we saw that this was not the case. Again, the real-world impact of the global pandemic is seen here and can be measured through further analysis to clearly measure how much deviation there was from the predicted numbers.

When comparing the model data to the existing data, we see not only that there was a deviation from the prediction, but that that deviation was in larger quantities of sales. While a marked increase in sales—and thus revenue—is likely to be a positive, this time series model is valuable in that it can aid in future planning for supply chain, distribution, and purchasing needs going forward.

# Technique and Tool Limitations

Though it was beneficial in this problem study, it is worth noting that the most impactful limitation of this model is the timing at which the data was gathered. ARIMA models struggle with adjusting for turning points in market need, which is clearly seen in the resulting comparison. This deviation will likely continue between the predicted model and the current sales model until new data is added to the dataset or the dataset is split between pre- and post-pandemic evaluations. As such, models must be rerun regularly and used largely for comparison, rather than as a goal definition.

# Summary of Proposed Actions

A sound recommendation is that the county spend time leading up to this point, putting in place processes and spending limits to lessen the impact of the revenue loss.

It is the researcher’s recommendation that as data continues to be gathered, the models be split into pre- and post-downturn evaluations and compared thusly. In doing so, the information will not only reflect overall growth (or loss) within the market, but also growth compared to the current economic truth that has impacted the county’s sales numbers. By evaluating the sales data as two separate entities—when there is enough data to support this shift—the county will have more accurate forecasts for decisionmakers to plan with.

Furthermore, in order to better analysis data in the future, it would be beneficial for the county to record data on sales per day, rather than per month. It would also be beneficial for the county to record gains and losses for the missing dates so that the standard deviation information could be replaced with reality.

Lastly, this evaluation be run on a quarterly basis to continue to reap the benefits of forecasting.

# Benefits of the Study

By implementing stronger data collecting practices, the county will see more accurate and responsive modeling immediately.

In running this analysis quarterly—with updated data parameters—the county will be able to measure their pre- and post-pandemic growth and loss accurately. This will allow decision-makers to adjust to the market’s “new normal” while maintaining information on their overall growth and potential downturn.

Accurate forecast models will also alleviate some of the global market’s supply chain instability by allowing stakeholders to make decisions on distribution and demand predictions.

**References**

Longo, A., & Fischer, J. (2020, December 22). *Montgomery County sees $11 million spike in alcohol sales in 2020*. WUSA9. https://www.wusa9.com/article/news/health/coronavirus/moco-abs-sales-2020/65-02b9cfc9-3da1-498c-894e-a08c4a1f4053

Tiao, G. (2015, April 10). Time Series: ARIMA Methods. International Encyclopedia of the Social & Behavioral Sciences (Second Edition). https://www.sciencedirect.com/topics/social-sciences/time-series