**Data Analytics Capstone Topic Approval Form**

**Student Name:** Nina S. Gooden

**Student ID:**  009823504

**Capstone Project Name:** Time Series Analysis of Retail and Warehouse Sales

**Project Topic**: Predictive Model for Retail and Warehouse Sales

**This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** Can an Arima time series model be created from the dataset in order to accurately predict next quarter sales for Montgomery County?

**Hypothesis**: H0: A time series model with accuracy > 70% cannot be made from the Warehouse and Retail Sales dataset.

H1: A time series model with accuracy > 70% can be made from the Warehouse and Retail Sales dataset.

**Context:** The contribution of this study to the field of Data Analytics and the MSDA program is to create a time series model which will predict product sales within the county. Montgomery County of Maryland collects data on product suppliers, as well as volume of sales for tax purposes. This study will create a forecast that the county can use to better understand the underlying costs, trends, or systemic sales patterns over time (Hayes, 2022). This study will utilize autoregressive integrated moving average (ARIMA) to describe autocorrelations in the data, while also using exploratory data techniques to derive general insights about the dataset. Lastly, the researchers will package findings in a digital dashboard for general consumption. It is the researcher’s hypothesis that there will be a clear prediction derived from this dataset.

**Data:** The data needed to be collected for the question is the publicly available information provided by the Montgomery Country of Maryland local government ESB Service *Warehouse and Retail Sales - Catalog*. (n.d.). The data was last updated December 9, 2022 and is updated monthly. There are 307,646 records in the data set and 9 columns.

The initial data set is made available through the data.gov website. The data set includes the following variables of year, month, supplier, item code, item description, item type, retail sales, retail transfers, and warehouse sales. The predictor variables are broken down as follows:

<https://catalog.data.gov/dataset/warehouse-and-retail-sales>

Graphical user interface, application, table, Excel

Description automatically generated

The updated data set used in this model was cleaned and truncated to answer the research question. That cleaned data set can be found here:

[OneDrive Cleaned Data](https://1drv.ms/x/s!AqOi3D28HmhEhZt_PGNORoEgNa7xgA?e=ERZfgj)

And maintains the following predictor variables:

|  |  |
| --- | --- |
| **Field** | **Type** |
| year | Categorical |
| month | Categorical |
| day | Categorical |
| item\_type | Categorical |
| total\_sales | Continuous |
| date | Categorical |

The data is intended for public access and use, and no license information is provided. The dataset is updated regularly and has been consistently managed since 2017, meaning the data available is healthy and of sufficient size. Forecasting a time series is fundamental for business planning, procurement, and production. Ensuring data is healthy is necessary, as any errors in forecasts will ripple down any business context (Prabhakaran, 2022). There will be no information in the final evaluation that would make the companies listed in these transactions identifiable. Limitations: The information available through public forums does not include price data or revenue information. As such, the time series evaluation will need to focus on number of sales, rather than their value. Delimitations: There are no limitations for the cleaned dataset.

**Data Gathering:** Any entries that have missing inputs for total sales will me imputed through data averages. Missing data in the item\_type field will be manually evaluated and inputted, which is the ideal substitution method for easily located, low-cost data (Mesidor, 2021). Records missing date information will be removed, as time series models work with the complete data and require impute of the missing values prior to modeling (Kumar, 2022). No other columns will be evaluated in this analysis. The data quality is remarkably high, as it has been collected by local government and maintained for five years. Data will be made stationary if it is not initially and correlation and auto correlation will be plotted visually. Overall data sparsity is < 5%.

**Data Analytics Tools and Techniques**: 1. Exploratory visualizations will be created for cleaned data to gather insights into which item type contributes what percentage of sales. 2. Initial visualization will be created using pyplot. 3. Stationarity will be evaluated and corrected if necessary, using autocorrelation function (ACF) and partial autocorrelation function (PACF) to find the p and q values for ARIMA. (Zvornicanin, 2022) 4. A portion of the data will be segmented for testing and assessment of the model. 7. ARIMA Model will be constructed. 5. Model accuracy will be measured with held-over data, as well as mean square root evaluation. 6. Initial visualizations will be created. 7. Data will be exported in order to create dashboard visualizations in Tableau. The auto\_arima and ARIMA functions will be used to create forecast data, though it is important to note that ARIMA has difficult prediction turning points. (*Understanding ARIMA Models for Machine Learning*. (n.d.)). This available data ends in 2020, which was a mark of a global turning point and thus the prediction model will likely exceed the reality of data.

**Justification of Tools/Techniques:** Python will be used for the duration of the project. According to (Manokhin, 2022), “when it comes to forecasting and time series in 2022, Python is the no-brainer choice.” In addition, (Alam, 2021) states that ARIMA is “arguably the most popular and widely used statistical technique for forecasting.” These sentiments justify a solid foundation in both for time series analysis that is accurate, reproducible, and scalable.

**Project Outcomes**: The project will seek to create a forecast model for the number of sales in the next quarter for the combined efforts of warehouse and retail for Montgomery County of Maryland. Support for the alternative hypothesis is found in *Time series analysis: validating effect of changes* with the statement “time series hypothesis testing talks about how we identify whether different time periods have significantly different observation” (Tunggawan, 2018).

**Projected Project End Date**: 12/31/2022

**Sources**:

Alam, M. (2021, December 15). Comparing the performance of forecasting models: Holt-Winters vs ARIMA. Medium. <https://towardsdatascience.com/comparing-the-performance-of-forecasting-models-holt-winters-vs-arima-e226af99205f>

Hayes, A. (2022, June 13). What Is a Time Series and How Is It Used to Analyze Data? Investopedia. <https://www.investopedia.com/terms/t/timeseries.asp>

Kumar, S. (2022, April 30). *4 Techniques to Handle Missing values in Time Series Data*. Medium. <https://towardsdatascience.com/4-techniques-to-handle-missing-values-in-time-series-data-c3568589b5a8>

Manokhin, V., PhD. (2022, December 16). Python vs R for time-series forecasting - Valery Manokhin, PhD, MBA, CQF. Medium. <https://valeman.medium.com/python-vs-r-for-time-series-forecasting-395390432598>

Mesidor, F. (2021, December 11). A basic guide to time series analysis - Towards Data Science. Medium. <https://towardsdatascience.com/a-basic-guide-into-time-series-analysis-2ad1979c7438>

Prabhakaran, S. (2022, September 3). ARIMA Model – Complete Guide to Time Series Forecasting in Python. Machine Learning Plus. <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>

Time Series Analysis: Definition, Types, Techniques, and When It’s Used. (n.d.). Tableau. <https://www.tableau.com/learn/articles/time-series-analysis>

Tunggawan, E. (2018, October 27). Time series analysis: validating effect of changes. Elvyna Tunggawan. <https://elvyna.github.io/2018/time-series-hypothesis-testing/>

*Understanding ARIMA Models for Machine Learning*. (n.d.). Capital One. <https://www.capitalone.com/tech/machine-learning/understanding-arima-models/>

*Warehouse and Retail Sales - Catalog*. (n.d.). <https://catalog.data.gov/dataset/warehouse-and-retail-sales>

Zvornicanin, E. (2022, November 8). Choosing the best q and p from ACF and PACF plots in ARMA-type modeling. Baeldung. <https://www.baeldung.com/cs/acf-pacf-plots-arma-modeling>

To be filled out by a course mentor:

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Mentor’s Approval Status: Select one

Date: Click here to enter a date.

Reviewed by: Click here to enter text.

Comments: Click here to enter text.