# Data Analytics Capstone Topic Approval Form

**Student Name:** Jessica Hosey
**Student ID:** 005944101

**Capstone Project Name:** Demand Forecasting Using Time Series Analysis to Optimize Inventory

**Project Topic**: Demand Forecasting using ARIMA

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

**Research Question:** To what extent can product demand be predicted using time series analysis of historical sales data?

**Hypothesis**: **Null Hypothesis**- Product demand cannot be predicted with 90% accuracy.
**Alternate Hypothesis**- Product demand can be predicted with 90% accuracy.

**Context:** The goal is to analyze the historical transactional data to forecast product demands for the company. Predictions on inventory would help drive business decisions about inventory management, marketing campaigns, and procurement.

**Data:** Historical transactional data is needed to perform the time series analysis to predict demand.

The project will use the Online Retail dataset from the UCI Machine Learning Repository. This dataset contains transactional data from a UK-based online retailer between December 2010 and December 2011. The following information is the columns and their descriptions of the dataset:

- InvoiceNo: Unique transaction number.
- StockCode: Unique product number.
- Description: Name of the product.
- Quantity: Number of units sold within a transaction.
- InvoiceDate: Date and time of transaction.
- UnitPrice: Price per unit of product.
- CustomerID: Unique customer number or identification code.
- Country: Customer's country location.

The dataset that is being used is owned by the UCI Machine Learning Repository. The Repository is a collection of datasets that can be used by the data science and machine learning community. We can use this dataset as it is licensed under Creative Commons Attribution 4.0 International, which allows it to be shared and used for any purpose as long as credit is given. Since this dataset is within the UCI Machine Learning Repository, it is available for anyone to use if the user cites the dataset owner.

**Data Gathering:** The Online Retail Dataset will be sourced from the UCI Repository. This will be the primary data source for this capstone project. The dataset contains 282,959 transactions from a UK-based e-commerce store. The transactional data includes the following columns: InvoiceNo, StockCode, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. The dataset is easily downloadable from the UCI Repository website. The dataset will then be uploaded to Jupyter Notebook, and data quality will be analyzed. Data cleaning techniques that will be analyzed are as such: checking for significant errors, missing values, date and time corrections, and incorrect product quantities (i.e., Additional information that may be helpful in the analysis will be gathered, such as the UK Holiday calendar. Once the data is loaded and cleaned, data aggregation techniques will be used to group sales data by day, week, and month. In addition, total revenue, moving averages, and total number of sales will also be calculated. Then, EDA will be performed to confirm and validate that the dataset captures the necessary patterns and has enough historical data to support forecasting analysis.

**Data Analytics Tools and Techniques**:
Tools:

- Jupyter Notebook: Anaconda Cloud
- Python: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Statsmodels

- Data Visualization: Tableau

Techniques:

- Data Preprocessing
  - Data Cleaning: Handle missing values and outliers and correct datetime data.
  - Data Aggregation: Aggregate sales data by daily, weekly, and monthly demand.
  - Model Feature Engineering: Rolling averages, lag features, etc.
  - Generate total revenue per product and moving averages
- Exploratory Data Analysis (EDA)
  - Trend and Seasonality: Identify seasonality, trends, and outliers in aggregate sales data. Time series data using line plots and decompositions to inspect trend, seasonal, and residual components.
  - Visualization: Use line charts and other diagrams to understand product demands, fluctuations, and outliers.
- Time Series Forecasting Model
  - ARIMA will forecast product demand for non-seasonal products. We could also use SARIMA for seasonal products.
- Model Evaluation
  - Mean Absolute Percentage Error to analyze model accuracy.
- Dashboard Engineering and Deployment: Tableau
  - Create an interactive dashboard with instructions about deployment. Included in the dashboard is:
    - Forecasted and actual demand for non-seasonal products.
    - Trends and seasonality visualizations
    - Key performance metrics and actionable business insights.

**Justification of Tools/Techniques:**

Tools:

- Python: A programming language that is effective in analyzing data, completing time series modeling, and creating visualizations.
- Pandas: Library used for data cleaning, aggregation, and manipulation.
- NumPy: Used to complete numerical computations.
- Matplotlib: Library used to create static visualizations such as line graphs.
- Seaborn: Library is used for more detailed statistical plots such as heat maps.
- Scikit-learn: Library used for feature engineering (lag features) and baseline models such as moving averages. This library also calculates model performance metrics such as mean absolute percentage error.
- Statsmodels: Library used for building ARIMA and SARIMA forecasting models.

Techniques:

- Data Preprocessing
  - Raw data contains many mistakes, missing values, inconsistencies, and outliers. Data cleaning ensures that the dataset is complete, accurate, and reliable.
  - Data aggregation allows us to capture patterns in demand for each product and reduces noise in the time series model.
- Exploratory Data Analysis (EDA)
  - EDA is important for understanding data trends, seasonality, and outliers necessary to build accurate demand forecasts.
  - Visualizations provide more insight into patterns and help us identify low or high product demand periods.
- Time Series Forecasting Model
  - ARIMA/SARIMA is used for classical statistical forecasting.
- Model Evaluation
  - Mean Absolute Percentage Error to quantify the effectiveness of the model's predictions. This metric allows us to compare models before selecting the best one.
- Dashboard Engineering and Deployment: Tableau
  - Used to build interactive dashboards to communicate business insights and actionable recommendations for stakeholders. Tableau allows the non-technical user to interact with the data to make informed business decisions.

**Project Outcomes**:

Outcomes:

1. Accurate demand forecasts for individual products and categories.
2. Insights into seasonal trends and key demand drivers.
3. Actionable recommendations for inventory and procurement planning.

Deliverables:

1. A dashboard for real-time data monitoring and business insights used for decision-making.

**Projected Project End Date**: 2/16/2025

**Sources**:

Chen, D. (2015). Online Retail [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5BW33.

**Course Instructor Signature/Date:**

☐ The research is exempt from an IRB Review.

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

Course Instructor's Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.