CS Rubric – Case Study Recreation

DS 4002 – Fall 2024 - Elaine Shagdarjav
Due: TBD
Submission format:

- **Upload GitHub repository link and pdf of Written Portion.**

Individual Assignment
General Description: Upload GitHub repository link and pdf of Written Portion.

Why am I doing this? The goal of this assignment is to familiarize you with time series data and ARIMA. Case studies are a great way to gain hands-on experience in dealing with how real world data, like the given datasets on immigration, GDP, and unemployment, interact with each other. By engaging in this case study, you will develop a deeper understanding in how to analyze and interpret time series data with ARIMA. This experience will also enhance your ability to draw meaningful conclusions from multivariate data. Additionally, reproducing a case study can help you learn what aspects of a study you want to employ in your future research, and what aspects you want to avoid. You will be able to apply both technical and analytical aspects of this case study to your life.

- Course Learning Objective: Bridge your knowledge with practical application, aiding in preparation of future challenges in the area of Data Science

What am I going to do? You will use the GitHub repository's datasets and code to perform ARIMA (Auto-Regresive Integrated Moving Average) modeling and cross-correlations. You will gain practical experience in applying advanced statistical techniques to real-world data, when to use said technique, and how to determine its parameters. You will learn how to interpret the results of the ARIMA model, understanding its strengths and weaknesses in the context of comparison. Additionally, you will explore the implications of your findings, considering what parts of the analysis plan were strong, and where it could be improved. Overall, you will find how immigration affects economic factors. Deliverables include:

- GitHub Repository - contains code and other necessary files
- Written Portion - as a PDF

Tips for success:

- Make sure you understand the main analysis code file. This case study provides you with base code that will help you create and employ the models you need.
- If you don't understand a line of code, try messing with it (changing parameters, moving its location, etc.) to see how its outputs change to figure out its purpose.
- If a certain aspect of ARIMA isn't making sense, refer to resources available online.

How will I know I have Succeeded? You will meet expectations on this Case Study when you follow the criteria in the rubric below.

Spec Category	Spec Details

Formatting	● GitHub repository (submitted via link on Canvas) ○ README.md ○ Code Folder ○ Output Folder ○ REFERENCES.md ● Written Portion ○ Submit as a PDF.
Written Portion	Discuss the results of your replication of this study: ● In a paragraph, summarize the motivation of the study and the goal ● In a paragraph, discuss your results and potential obstacles to accuracy Reflection: ● In one or two paragraphs: ○ Discuss one or more technical challenges you experienced while completing this case study and how you solved them. ● In one paragraph, describe one analytical/research aspect from this experience you want to incorporate into your future projects, and describe one where you would not
GitHub Repository	README.md ● This will include a summary on what you have produced for the case study. Does not need to be detailed but should provide a high-level overview of what you have done in the repository. Code Folder: ● Well-documented Jupyter Notebook file(s) (.ipynb) should include code that uses cs_dataset.csv and: ○ Performs an exploratory data analysis of these factors and plots raw data: ■ immigration rates as a percentage of the US population (this variable must be made) ■ GDP as a percent change (GDP_PC1) ■ Unemployment rates ○ Determines ARIMA parameters through: ■ Augmented Dickey-Fuller Test ■ Analyzing ACF/PACF graphs ■ Analyzing raw plots made in previous step to determine stationarity ○ Employs ARIMA on each variable in cs_dataset.csv and create residual graph and residual density graph ○ Performs cross-correlation of already ARIMA-employed data with time lags and outputs results Output Folder: ● Plots created from source code files ● .txt of cross-correlation results REFERENCES.md ● Markdown file with citations to any resources referenced in your recreation of this case study in an IEEE documentation style.