Case Study Rubric – Predicting Taxi Data

DS 4002 – Instructor: TBD
Due: TBD
Submission format: link to GitHub repository submitted on Canvas

General Description: Submit to Canvas a link to your GitHub repository for this case study

Why am I doing this?
This case study provides the opportunity to use the skills and technique you have used in other Data Science Courses for the analysis of real-world data. Specifically, this case study will focus on conducting time series data analysis and building predictive models. You will gain a greater understanding of how to work with time series data and how to employ various evaluation metrics that situate the results you reach.

What am I going to do?
This case study entail retrieving data from the Chicago Data Portal and cleaning and preparing the data for time series analysis or utilizing the cleaned data provided. You will then create EDA plots to visualize any trends within the time series data and to justify which models to use for data within the month of October. Follow this, a SARIMA and BSTS model should be built to predict a day in October and the performance of each model for this prediction will be evaluated. Deliverables include:

- GitHub Repository – This will contain all the project materials (Data, models, EDA graphs, code)
- One page document – A PDF reflecting on the process of completing this case study

Tips for success:
- Spend some time learning more about time series data. This will help frame the analysis you will conduct as you work through this case study
- Learn (or refresh) yourself on how R works as this is the language originally utilized to complete this project. Doing this earlier on helps streamline the shift to working with R
    - This includes visualization!
- Reference additional resources to fully understand what makes up a SARIMA and BSTS model. By understanding how each model is built and the different components they entail, you will be able to grasp how each model works and what the results mean

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 Link to your repository**
	  ○ Create your own GitHub repository titled "CS2Taxi-SEMYEAR" (ex. "CS2Taxi -F25", for fall 2025 semester) that contains:
	  ▪ README.md file
	  ▪ LICENSE.md file
	  ▪ REFLECTION file
	  ▪ SCRIPTS folder
	  ▪ DATA folder

	▪ OUTPUT folder o Submit this link to Canvas • **Written Section** o One page, PDF o **Uploaded to GitHub Repository**
GitHub Repository	• **README.md file** o **Goal:** Provide a quick outline of what this repository contains o Describe the software and platforms that you used for this case study ▪ Include any add on packages utilized within R o Provide a description of how this repository is organized (note the hierarchy of folders, subfolders, and which files are in each) • **LICENSE.md file** o **Goal:** Describe to those who visit this repository the terms in which visitors can utilize and cite this repository o Often a MIT license is sufficient, but look into the GitHub options to make sure this is the one you want • **SCRIPTS folder** o **Goal:** This folder should contain all the code and scripts employed for this case study o This folder should NOT CONTAIN any other files that do not contain source code o Make sure to name each script with a descriptive title that reflects which step(s) that code seeks to complete • **DATA folder** o **Goal:** This folder holds all the data utilized for this case study o The Taxi data (cleaned and uncleaned) should be stored here o Include a data appendix file that details: ▪ What a unit of observation is ▪ Variables included within the Chicago Taxi dataset ▪ Summary statistics for each variable (where applicable) ▪ A Graph of the Taxi Ridership count across October 2023 • **OUTPUT folder** o **Goal:** This folder should hold any graphs, plots, figures, tables, etc. that come from running the code o Give each file a descriptive name that reflects what that table/graph/plot/etc. is
Code	• When conducting this case study, there should be a clearly labeled script for: o Obtaining and Cleaning the Chicago Taxi Data o EDA plots for looking at taxi rideshare trends

	○ Building the SARIMA model and Predicting October 30th, 2023 ○ Building the BSTS model and Predicting October 30th, 2023 ○ Analysis of Evaluation metrics (RMSE, MAE, MAPE) • All code scripts should have descriptive comments through them detailing what each line or chunk of code does
Written Section	• **Goal:** Reflect and gain insight into how this Case Study went for you ○ Provide a short executive summary, outlining what this document will talk about ○ In one paragraph, describe two aspect you liked or enjoyed while working through this case study ▪ If you didn't like anything, then say why! The main question is: how is this informing your interest? ○ In one paragraph, describe one way the results from this case study will inform something in your daily life pertaining to rideshares • **Upload this PDF to the GitHub Repository**
References	• All references should be listed at the end of the written section PDF • Use IEEE Documentation style (link)

Acknowledgements: Thank you Professor Alonzi for the rubric outline!!!