Divya Lobo
DS4002
CS3
12/9/24
Professor Gates

CS1 Rubric – Predicting Uber Prices in New York

DS4002- Fall 2024- Professor Gates

Submission Format: Github repo

Individual Assignment

General Description: In this case, you will analyze whether Lyft average total fares can be used as a factor in predicting monthly Uber average total fares. To do this, you will obtain ride fare data from the NYC Taxi and Limousine Commission, and calculate the average total ride fares per month. You will compare the mean absolute error of SARIMA models that predict the monthly average uber fares with and without Lyft fares as a covariable.

Preparatory Assignments: Project 2 Milestones

Why am I doing this? This project allows you to explore SARIMA models and forecasting. You will also get to learn how to handle large data files like PARQUET files. You will get to explore New York Uber and Lyft fares between 2019-2024 and identify any trends over time.

What am I going to do? You are going to create a simplified data frame of Uber and Lyft rides from source data consisting of average fares from every month between September 2019- July 2024. You will create exploratory graphs of Uber and Lyft fare distributions over time and in comparison to each other. Finally, you will create an SARIMA model to forecast the average Uber price for each month and an SARIMA model to forecast the average Uber price for each month using Lyft as a covariable. You will compare the mean absolute error to determine if Lyft can be used to predict the average Uber fare in New York each month. Deliverables include a github repo containing:

- The simplified dataset of Uber and Lyft fares
- Exploratory graphs of Uber and Lyft rates over time and in comparison to each other
- Scripts for cleaning the data and creating the data exploration and SARIMA models
- A ReadMe file explaining how to run the repo for recreation

How will I know I have Succeeded? You will meet expectations on CS1 when you follow the criteria in the rubric below.

Divya Lobo
DS4002
CS3
12/9/24
Professor Gates

Category	**Spec Details**
Formatting	One Github Repository (submitted via link on canvas) ● To ensure reproducibility, the repository will adapt parts of the TIER Protocol 4.0. In a nutshell, the top level page of the repository should contain: ○ A README.md file ○ A LICENSE.md file (use MIT as default) ○ A SCRIPTS folder ○ A DATA folder ○ AN OUTPUT folder
README	Goal: This file serves as an orientation to everyone who comes to your repository, it should enable them to get their bearings. ● Use markdown headers to divide content ● Make an H2 (##) section explaining the contents of the repository ● Provide Instructions for reproducing your results: ○ In this section, you should give explicit step-by-step instructions to reproduce the Results of your study. These instructions should be written in straightforward plain English, but they must be concise, but detailed and precise enough, to make it possible for an interested user to reproduce your results without much difficulty.
SCRIPTS	Goal: This folder contains all the source code for your project. ● Include all the scripts you used. Try to name each script according to the order it needs to be executed to reproduce the results. ● There should be a minimum of two scripts: cleaning the data and data analysis ○ The cleaning the data script should output a csv file that can be used in the data analysis script ○ The data analysis script should include the process of creating data exploration graphs, the ARIMA models, validation for the models, and plot the forecasts for the models with and without Lyft fares as a covariable ● All script files should include header comments at the beginning of a script to provide information that anyone working with or executing the script should be aware of. Throughout all your scripts, you should include comments explaining what each command or sequence of commands accomplishes and what the purpose is.
DATA	Goal: This folder contains all of the data for this project. ● Provide a link to the NYC Taxi and Limousine Commission data set or a cloud storage file containing all the PARQUET files from the source data. ● Provide the final csv that the cleaning data script returns. ○ This csv file should contain a column for the date (month and year), the average Uber total fare, average Lyft total fare, average Uber base fare and average Lyft base fare. ● A Data Appendix file as a PDF, which will include tables, figures, and other descriptive statistics. This file should be organized in sections. After that, you should include a subsection for each variable in the analyzed dataset.

Divya Lobo
DS4002
CS3
12/9/24
Professor Gates

	○ More information: https://www.projecttier.org/tier-protocol/protocol-4-0/root/data/analysisdata/data-appendixfile
OUTPUT	Goal: This folder contains all of the output generated by your project, e.g. figures, tables, etc. ● Use informative names for your files
REFERENCES	All references should be listed at the end of the document ● Use IEEE Documentation style

Acknowledgements: Special thanks to Pete Alonzi, Javier Rasero, and Jess Taggart from UVA CTE for coaching on making this rubric. This structure is pulled from Streifer & Palmer (2020)