CS3 Rubric – Case Study

DS 4002 – Catherine Nguyen
Due: TBD

Submission format:

- Upload link to GitHub repo to Canvas

Individual Assignment

Why am I doing this?
This case study is your opportunity to utilize data science skills and tools to analyze trends and patterns, run statistical testing, and deliver data driven insights. This assignment will allow you to practice working with text data and sentiment analysis in a real-world context.

What am I going to do?
The GitHub repository for this case study can be found at: https://github.com/Catherine-Nguyen88/DS4002-CaseStudy/tree/main. You will have a cleaned dataset of 20 TV shows of various genres. You will use Python and a platform of your choice (e.g. Google Colab) to run EDA, VADER sentiment analysis, and multiple linear regression (MLR) on the cleaned dataset. The resulting EDA plots and MLR statistical test results will indicate which of the four sentiment (positive, negative, neutral, compound) are significant and for what types of shows. Deliverables include:

- A data dictionary of important variables, data types, and their definitions
- A Python script used for EDA and analysis (well-documented)
- A document of plots and statistical test results
- A GitHub repository containing all data files, scripts, and documents used

Tips for success:

- Don't overthink it. You are following a case study with all the resources you need readily available to you.
- Talk to your fellow students. If you don't know how to complete a step, consult a peer. They can offer assistance.
- Observe the multiple linear regression statistical results such as p-value, R-squared, variable coefficients and F-statistic to help form your conclusions.
- Work in Python, if possible. The instructions laid out will be easier to follow as they were written in this language.
- Know exactly which folder has what resources you need before starting the case study (refer to README in GitHub repo to orient yourself to the important folders in the repo).

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	• Repository – A GitHub repo (and cloud storage folder if necessary) containing all materials
	  ○ Submit a link to the repo, title the repo 'CS3_SentimentAnalysis_ShowRuntime'
	  ○ Contents
	  ▪ README.md
	  ▪ LICENSE.md
	  ▪ A DATA folder
	  ▪ A SCRIPTS/OUTPUT folder
	  ○ Use pdf format when possible

README.md	Goal: This file serves as an orientation to everyone who comes to your repository, it should enable them to get their bearings. • Include a map of your documentation: In this section, you should provide an outline or tree illustrating the hierarchy of folders and subfolders contained in your Project Folder, and listing the files stored in each folder or subfolder
LICENSE.md	Goal: This file explains to a visitor the terms under which they may use and cite your repository. • Select an appropriate license from the GitHub options list on repository creation. • Usually, the MIT license is appropriate.
DATA folder	Goal: This folder contains all the data for this project. • You should include the initial data, and the final data analyzed. • Initial data: 20 TV script pdfs • Final data: Sentiment scores excel produced from source code file. Call this 'all_sentiment_scores.xlsx'.
SCRIPTS/OUTPUT folder	Goal: This folder contains all of the code scripts and output generated by your project, e.g. figures, tables, etc. • Source code file - Python script (EDA and analysis) called 'CS3_script' • PDF document with data dictionary • PDF document with all plots and statistical testing results

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