Keith Lee Case Study Rubric

DS 4002 – Fall 2024 - MaryGrace Gozzi
Due: December 9th, 2024
Submission format: Upload link to GitHub repository via class Canvas page

Please Note: This is an individual assignment.

Why am I Doing This? As you work through this case study, you will gain experience with data cleaning, sentiment analysis, and creating data visualization. Fundamental skills, like naming variables and keeping data frames organized, are vital to any data science endeavors. More advanced skills, like implementing sentiment analysis, allow for more nuanced conclusions and patterns to be identified in data. Finally, making effective graphical representations to explain your findings is an important step in the project process. Working on these skills will leave you more prepared for a career in data science, or future academic pursuits.

What am I Going to Do? You will begin by doing some reading on 6 restaurants, each reviewed by popular food influencer Keith Lee, to determine if he left a positive or negative review. You should also make a note of the date that Lee posted his restaurant review. Next, you will scrape Yelp.com, a publicly available website, to collect customer-written food reviews. The reviews should be sorted into 'Before' and 'After' sections based on the posted date of Lee's review that you noted before. Then, you will write code to determine the 'sentiment' of each review- consider using the VADER sentiment analysis tool. Once you have determined the sentiment of the reviews (positive, negative, or neutral), you should find the ratio of positive to negative reviews in both sections. Finally, you will find the change in ratio between the 'Before' and 'After' sections. That information will allow you to determine: did Lee's positive reviews increase the percentage of positive reviews? Did his negative reviews increase the percentage of negative reviews?

Deliverables include:

- A document outlining the following: Project Goal Statement, Research Question, and Hypothesis
- Code scripts used to determine sentiment of food reviews, as well as the ratios of positive and negative reviews for each restaurant ('Before' and 'After')
- A GitHub repository that contains data for all the restaurants, code scripts, and instructions for project reproducibility.

Tips for Success:

- Do not let your own assumptions or biases cloud your interpretation of the data- no matter what you may expect to happen given Keith Lee's reviews, focus on the story the data is telling. Avoid making assumptions!
- Keep your data organized. There are 6 restaurants involved in the study, and reviews are being considered in 'Before' and 'After' time periods. and it is important to make sure your restaurant reviews are saved and analyzed separately.

    - Best Practice: Pick one naming convention for your CSV and Python files. For example, all files related to 'Before' reviews could be 'Before_RestaurantName_Reviews'.
- Do not be afraid to ask questions! Reach out to your professors during office hours, and bounce ideas off of your peers.

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

Formatting	● One GitHub repository, submitted via link on Canvas ○ Repository should be named 'Keith_Lee_CS_LastName' ● Content of the repository should include: ○ a README.md file ○ a Data Folder ○ an Output folder ○ a REFERENCES.md file
README.md	● Goal: This file serves as an orientation to everyone who comes to understand your work. ● Sections should be organized in the following way: ○ *Section 1:* Software and platform section ▪ Include: The name of platform used to conduct analysis, and the name of the software used for the project ○ *Section 2:* An outline of your project documentation ○ *Section 3:* Specific instructions for reproducing your results, listed with enough detail that someone could follow them without any outside guidance
Data Folder	● Goal: This folder will contain all of the data for this case study ○ This includes the data include the initial data as well as the final data ○ Also include a Data Appendix with the following items: ▪ text (typed explanations) ▪ tables and figures ▪ any other relevant descriptive statistics
Scripts Folder	● Goal: The folder will contain all code files (.ipynb, .md, etc.) used to execute the goal of the case study ○ All code should be thoroughly commented to ensure flow and reader understanding

Output Folder	● Goal: This folder contains all of the output generated by your case study. This will include: ○ Figures ○ Tables
REFERENCES.md	● Include a file with all references used throughout the case study ○ This may include: journal articles, websites, or written sources ● All sources should be cited with IEEE Documentation style