California Restaurant Reviews Case Study Rubric

DS4002 - CS 3 - Mohini Gupta
Individual Assignment

Why am I doing this? This is your opportunity to apply some of your data science tools and skill sets to solve a problem. The materials and scenario are sourced from a real-life situation, and you will get to explore the implications of real world data in a controlled setting.

- Learning objective: understand how to use data science skills to solve real world problems

What am I going to do? Now that you have some experience with creating models, visualizations, and communicating results, you're going to put those skills to the test in this case study. You will be analyzing text data, in the form of reviews acquired from Yelp's open dataset, and creating a model to find out which words are most strongly associated with low-rated and high-rated reviews. You will also find out what the sentiment scores are for each rating level. Then, you will go over your results and interpret them. Remember, in this case study, you are playing the part of someone who is traveling across California and trying new restaurants. Be creative! Finally, you will be reviewing this case study and outlining what you learned through it. Deliverables include:

- GitHub repository: forked from the one provided in this case study and contains your findings, analysis, and reflection (https://github.com/rde6mn/DS4002-CS3)
- Keywords for low and high-rated reviews
- Review sentiment scores
- Analysis of results from the assigned role in the case study - what do your results mean in context of the scenario?
- Reflection - highlight the skills that you gained from doing this case study

Tips for Success:

- Be patient. Yelp's open dataset is very large and due to the complex nature of the analysis, some chunks of code may require up to an hour to run
- Start early. The earlier you start working on this case study, the less stressed you will be about having to complete all the parts.
- Ask for help. If you are struggling, don't hesitate to ask a friend or reach out to your TA or professor!
- Have fun! As mentioned before, some chunks of code take a long to run, so in the meanwhile, watch a movie or a show!

How will I know that 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	• Fork the provided GitHub repository for this assignment
	  ○ Everything is contained within the repository
	  ○ Contents:

	■ README.md ■ LICENSE.md ● MIT license ■ SCRIPTS folder ■ OUTPUT folder ■ Analysis ■ Reflection ○ Remove all other files that are in the forked repository
README.md	● 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. ● Section 1: Objective ○ What are you doing for this project? Why are you doing it? ● Section 2: 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. ● Section 3: Results ○ Summarize your findings. ○ If a person were to only look at your README, what would you want them to take away from your project?
LICENSE.md	● Goal: This file explains to a visitor the terms under which they may use and cite your repository. ● A license should already be provided from the original repository. If the license is deleted, use the MIT license.
SCRIPTS	● Goal: Contains code necessary to generate all outputs ● Organize all of your code for this case study into one folder ○ Code should be able to: ■ Clean and structure Yelp dataset ■ Run keyword analysis ■ Generate sentiment scores by rating level
OUTPUT	● Goal: Contains all outputs generated by your code for this case study ● Outputs should match up to what your scripts generate ● Must contain: ○ Top keywords for low and high-rated reviews ○ Sentiment score table
Analysis	● Goal: Understand implications of your outputs and summarize your findings ● PDF format, one page maximum
Reflection	● Goal: Summarize lessons learnt from this case study ● PDF format, one page maximum

Acknowledgments: Special thanks to Professor Alonzi for this rubric structure