[Janet Taylor](mailto:janett@g.clemson.edu) and Evan Kessler

Dr. Hubig

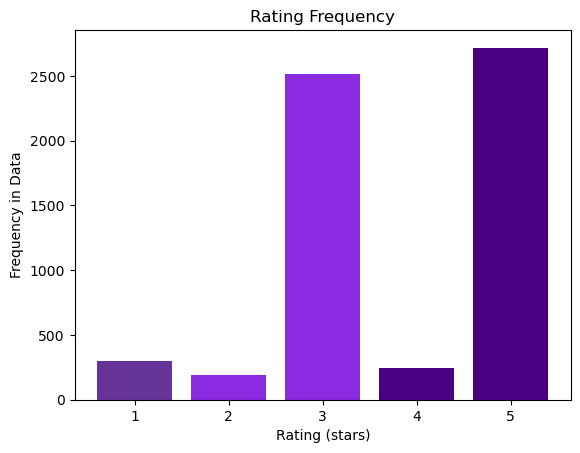
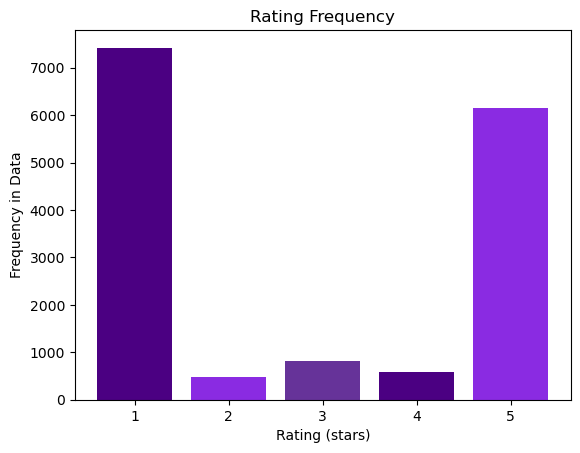
CPSC 4300: Applied Data Science

23 February 2023

Checkpoint 1

Unfair reviews can limit workers’ future job opportunities and can also result in workers not getting paid or even being terminated from the marketplace. Therefore, this semester we will be studying how negative reviews affect the gig economy (the part of industry where individuals pick up jobs that help share their resources i.e. driving people to places or picking up people’s food). Our units of analysis are the reviews and the ratings. In this project, we will be correlating what reviews people left, why they left those reviews, and if those reviews could be automatically detected to be out of control of the driver. We have 15,455 data points (filtered down to 5,972 data points) in our data set, with a range of reviews from 1 to 5 stars. Each review is a unique observation because each review refers to a single incident between a customer and their riding/driving experience. We have included data for Uber and Lyft, where we have two different datasets for Uber and one dataset for Lyft. We are using multiple different data sources to limit bias and to confirm our hypothesis against different websites. The data was taken from January 27, 2014, to February 19, 2023.

To gather our data, we signed up for a free premium trial on popular data scraping service, Octoparse, and parsed through trustpilot reviews using a custom automated workflow on the Lyft and Uber trustpilot webpages. We then exported the data into a CSV file and parsed the data using Pandas. Throughout collecting our data, we saw that there were numerous different types of customer review gripes. For instance, some reviews complain about the prices or about corporations which the actual drivers do not set. To clean the data, we undertook a series of steps. We stripped a few columns from the dataset that were likely to be completely irrelevant to the analysis (like the avatar of the user who left the review) or columns that were extraneously added by Octoparse. We did some filtering on the columns which had extra string text in them to extract the integer values we were interested in. Our data, consistent with how reviews are often given, was originally heavily weighted towards either 1 star or 5 star, but after filtering to reviews that explicitly mentioned drivers the ratio of ratings dramatically changed.



We will be using sentiment analysis to determine our outcomes. The key predictors we are going to explore are the sentiment of the reviews and the rating. We would like to explore these data points through different websites and the two different companies to correlate the word choices. We will use a regression model that can provide an output expressing the likelihood that the review is unfair or not. Ideally this output could be used by the company to decide whether the review should count against the driver’s record or not. We plan on doing bar graphs to visualize this word data with the correlation coefficients / weights, as well as bar graphs with the actual ratings compared to the word data. Finally we will plot the correlation coefficients with the actual ratings to prove the validity of our model and exclude any words that have a neutral sentiment. Another possible data set is our sentiment analysis altogether of ratings versus dates for the two respective companies, Lyft and Uber, to see which company is doing better. We may keep these two data points separate and perform sentiment analysis together and separate to see if the issues in the gig economy are pertaining to one specific company or if the issues are shared between both corporations.

Other predictors we are considering are when the review was placed, whether the company solicited the review or not. We will perform some analysis to determine if these factors affect the likelihood of unfair reviews as well. If so, they will be weighted in our regression model, and we will also provide charts for all predictors to explain why we decided to include them in the model.