Matt Raymond
Professor LaHaye
CPSC 392
10 September, 2019

Assignment 1

Provide a brief write up of the analysis

- How many instances are in the data?
    - *There are 4000 instances of data*
- How many attributes?
    - *There are 86 attributes in the whole file, but I'm not going that in-depth. I include 29 attributes in my main graph, but then I focus on 9.*
- Are the attributes numerical or categorical?
    - *These types are entirely categorical. Specific categories are formed, and then people are put into those different categories based on their zip code. Those numbers are turned into percentages of the population of the zip code, and then more or less rounded down to the nearest 10%. The specifics can be found in the documentation.*
- Any obvious outliers?
    - *There are a few obvious outliers (categories that are either extremely prevalent or extremely rare in a zip code), which can be seen in the plot of df1. However, I don't think that these are errors. Some zip codes are especially rich (like Beverly Hills), and some are especially poor (like Flint), resulting in extreme outliers.*
- Comment on any interesting trends you discover (describe the plots)
    - *One thing that I found pretty interesting is that the percentage of unskilled laborers in a zip code had a pretty strong correlation with the percentage of people with no car (in the correlation graph of df2). In the same graph, we can also see that there's a strong correlation between the percentage of entrepreneurs and farmers in a zip code, and the percentage of people in a zip code who have two cars. This isn't surprising on its own, but what did surprise me is that it's a stronger correlation than that between middle management and 2 cars. I think that this is probably because farmers and entrepreneurs need multiple cars for their jobs (like delivery/transportation), while middle management positions might tend to have "traditional" households where the husband works and the wife stays at home. As a result, only one car is needed. This requires more analysis though, and I wouldn't base any conclusions on it as of yet.*
- Include at least 5 different plots
    - *Included in the code*