Assignment 3: Exploratory Data Analysis in R

Problem 1

This problem involves the use of the 2017 National Household Travel Survey (NHTS) data to understand the travel behavior characteristics and test hypotheses regarding travel behavior. The data only includes the weekday trips for the Orlando-Kissimmee-Sanford adult subsample. Use the provided trip file (each record corresponds to one trip) to answer the following questions.

a. Provide descriptive statistics (minimum, maximum, mean, standard deviation) for continuous variables trip duration, trip distance, household size, household vehicle count, driver count, number of adults, number of workers and age.
b. Provide frequency distributions of categorical variables Gender, educational level, race, trip purpose, travel day and trip mode.
c. What is the frequency distribution of trips by purpose for the following market segments?
   I. Males vs Females
   II. Workers vs Non-workers
d. Based on the mode of transportation, what is the frequency distribution for the following trip types
   I. All trips (Male vs Female)
   II. Home-based Work (Male vs Female)
   III. Home-based Shopping (Male vs Female)

Problem 2

You are provided with a dataset titled “Florida County Data.csv” that includes the following variables:

1. RW: The typical ratio of the wage for a specific job (e.g. framing carpenters) in each county relative to the state average wage for that job in 2013.
2. PCI: 2012 per capita income.
3. POP: 2012 population.
4. WDEN: Weighted density. Population density, typically calculated as population divided by land area, is of limited usefulness as much land area is irrelevant, e.g. land in a national park can’t be developed. Instead, imagine asking residents how many people per square mile live near them, and taking the average of their answers. This would measure the average density perceived by residents. Weighted density measures this by calculating density at fine geographic levels and taking a resident weighted average.
5. SH65UP: The share of the 2012 population age 65 or older. This proxy the relative importance of in-migrant retirees in the local economy.
6. SHLH: The share of 2012 employment in the Leisure and Hospitality sector. This proxy the importance of tourism in the local economy.

The goal of this problem is to write a brief report to tell someone who knows nothing about FL, how relative wages vary across the state. Use R to do the following analyses prior to writing anything.

1. Calculate summary statistics for each variable.
2. Produce box and whisker plots and histograms for each variable.
3. Calculate summary statistics for each variable, weighted by county population.
4. Create new variables that equal the natural log of POP, WDEN, PCI, and RW. Name them lnPOP, lnWDEN, lnPCI, and lnRW.
5. Produce the correlation matrix for lnRW, lnPCI, lnPOP, lnWDEN, SH65UP, and SHLH.
6. Produce scatter plots for lnRW (vertical axis) against lnPCI, lnPOP, lnWDEN, SH65UP, and SHLH.

Please provide an intuitive explanation of all the tables and graphs included to answer the questions above. Please include your R code in the appendix or at the end of your document When submitting any work, your objective is to communicate information to the reader (in this case, your instructor) in a Clear, Concise, Complete, Careful, and Courteous manner (5 C's of good writing). If your work does not possess the "5C" qualities and/or does not adhere to the guidelines specified below, you will definitely lose A LOT OF credit even though you may have the correct answer(s).