**BSTT 413 – Introduction to Data Analysis with R**

**Homework 1**

**Due: Sunday, June 6th 2021 at 11 PM**

The paper by Smiley et al. (Nutrients, 2019) examines the relationship between sleep duration and metabolic syndrome (MetS) using the NHANES 2013/2014 data.

* The authors defined MetS using the criteria established by Lee, Gurka and DeBoer (Expert Rev. Cardiovasc. Ther, 2016), which takes into account racial and age-specific differences in populations, resulting in a complex formula to determine a score.

For this exercise, we will rely on a simpler criteria published in the joint scientific statement on metabolic syndrome (Alberti et al, 2009).

The criteria for a positive incidence of MetS is: “

1. elevated waist circumference (≥88 cm for women and ≥102 cm for men),
2. elevated triglycerides (≥150 mg/dL) or drug treatment for elevated triglycerides,
3. low HDL cholesterol (<40 mg/dL for men and <50 mg/dL for women) or drug treatment for low HDL cholesterol,
4. elevated blood pressure (systolic ≥130 mm Hg, or diastolic ≥85 mm Hg, or both) or antihypertensive drug treatment for a history of hypertension, and
5. elevated fasting glucose (≥100 mg/dL) or drug treatment for elevated glucose.” (Moore et al., 2017; *Preventing Chronic Disease*).

Additionally, Smiley et al. describe exclusion criteria to remove potential outliers, which must be applied. Specifically, subjects who reported sleep durations greater than mean + 2SD (two standard deviations) were excluded.

Use the above information to answer the following questions:

1. Re-create Table 1 in Smiley et al. using **data from the 2017/2019 cycle of NHANES** and relying on the MetS criteria described above. Your table, however, should also include summary statistics of the variables in the overall sample (prior to stratification by MetS status). It is not necessary to compute the p-values as in the article. (20 points)
2. Compare the proportion of MetS in the 2017/2018 data to the proportion reported in Smiley et al. If it is expected that MetS prevalence remains consistent across years, what can be said about the sensitivity/specificity of the criteria described above compared to the method used by the authors? (5 points)
3. Generally speaking, the mean and standard deviation are not recommended for highly (right) skewed variables. Such variables (e.g., income) tend to have a mean that is substantially larger than the median. Identify which of the numeric variables are likely right-skewed and report for them the appropriate measures of centrality and dispersion. (5 points)

It is recommended to use R Markdown to submit your assignment and the code used as a single document. You may, however, write your answer in a Word document and submit your code separately. In either case, make sure that your code is appropriately annotated using comments throughout.