CMPE – 428

**Assignment # 1**

**Exploration and Visualization of Data**

**• • • TASK 1**

1. Download the dataset “NHANES0910modified.csv” from <https://staff.emu.edu.tr/hakanaltincay/Documents/CMPE428/>. The columns named as ClassLabel, Gender, Race, Education, FamilyHistory and HighBloodSugar are categorical and the others are numerical.
2. In the dataset, there are some columns having missing values. Which variables have missing values? Impute the continuous values using mean and the categorical variables using the most frequent value. Report the values used for imputation.
3. Report the number of Positive and Negative samples in the dataset.
4. For how many Positive samples do we have Gender =1 and Race =2?
5. Which Race value is the most frequent in Negative samples?
6. What is the average BMI value of the samples having Education =3?
7. Compute the mean and standard deviation of BMI values, separately for Gender = 1 and Gender =2.

**• • • TASK 2**

1. For each continuous variable in the dataset, compute a pair of boxplots, one for each class (Positive or Negative) specified in the first column of the dataset. For WaistCircumference, it must look like as given below. Which variables have outliers?

![Chart, box and whisker chart

Description automatically generated]()

1. Provide a table of quartiles, Q1, Q2, Q3, and median values for each continuous variable in the dataset.
2. Report the number of outliers for each continuous variable.

**• • • TASK 3**

1. For each categorical variable in the dataset, compute the bar-plots for different categories. The bar-plots should present the number of samples having each particular value. For example, for Race, you should have a plot similar to the following:



**• • • TASK 4**

1. For each pair of continuous variables in the dataset, compute a scatter plot. For example, for BMI and WaistCircumference, you should have a plot similar to the following:

Chart, scatter chart

Description automatically generated