**Data Manipulation Case Study**

Consider the data set audit.csv, it has characteristics of 2000 tax returns, the data set includes the following variables:

**ID:** Unique Identifier for each person

**Age:** Age of person

**Employment:** Type of Employment

**Education:** Highest level of education

**Marital:** Current Marital Status

**Occupation:** Type of occupation

**Income:** Amount of Income declared

**Gender:** Gender of Person

**Deductions:** Total amount of expenses that a person claims in their financial statements

**Hours:** Average hours worked on a weekly basis

**RISK\_Adjustment:** The continuous target variable; this variable records the monetary amount of any adjustment to the person’s financial claims as a result of a productive audit. This variable is a measure of the size of the risk associated with the person.

**TARGET\_Adjusted:** The binary target variable for classification modeling (0/1), indicating nonproductive and productive audits, respectively. Productive audits are those that result in an adjustment being made to a client’s financial statement.

Q1. Find the average hours worked by:

(a) Each gender

(b) Each occupation category

(c) Marital Status

(d) Marital Status and Gender\*

Q2. How many people had zero deductions? Out of these how many had income greater than 70,000?

Q3.Compute the %age of productive audits across\*\*

(a) Gender

(b) Marital Status

(c) Education level

Q4: Find the quartiles for:

(a) Age

(b) Income