Pre-req: Download the following two files from:
https://github.com/vertica/DistributedR-demos/tree/master/census-shiny/dataset (Links to an external site.)

- “census-income.data.gz”. We will use this dataset as input for this homework. This dataset contains a sample population and their characteristics such as age and income. Each row has 42 columns.
- “census-income.features”. This file contains the names of the 41 attributes (column names) of the census-income.data file. For example, the first column in age. The 42’nd column name is “income”. Note that this file is missing the 42nd column name (we won’t use the column)

Task 1: Statistics about data (15 points)

Write a R function called “censusSummary(dataFile)” which performs the following tasks:

- Reads the input census data file whose path is given by “dataFile” and creates a data.frame.
- Prints the average age of all the people (column 1 depicts the age).
- Prints the number of females in the dataset (column 13 is the gender of the person)[Hint: your data.frame may store column 13 as a R “factor”]

Note: To obtain full score you should write each of the above tasks with as few R statements as possible. You should be able to solve each sub-tasks in a single line each (use R’s built in functions as much as possible).

Task 2: Database like queries on data.frames (10 points)

Use the “dplyr” package to solve the following tasks. Read about the “dplyr” package at:http://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html (Links to an external site.)

Write an R function called “dbSummary(dataFile)” which performs the following tasks:

- Reads the input census data file whose path is given by “dataFile” and creates a data.frame.
- Use “dplyr” functions to sub-select the group of people whose age is between 20 and 50, i.e, 20<age<50 (column 1 depicts age). Let’s call this group “G”. Print the average number of weeks worked by this group “G” (column 40 represents the number of weeks a person worked)
- Plot the ages of people in group “G”. The plot should show the age of people in ascending order (i.e., agea should be sorted). In your plot, the Y-axis should represent age, while the X-axis goes from 1 to N, where N is the total number of people in group “G”. Y-axis label should say “Age”.