Introduction to Data Science with R

Homework Assignment #2

HA #2 is analyzing the data file **income_2017.csv**. On top of variables you are already familiar with, the dataset includes the hourly wage and monthly hours for salaried employees, and their economic branch. List of economic branches can be found at: https://www.cbs.gov.il/en/publications/Pages/2015/Standard-Industrial-Classification-of-All-Economic-Activities-2011-Updated-edition.aspx

Please read the instructions carefully before solving the assignment:

- Create an R script that performs the required operations and submit the **R script file only**.
- Do not use absolute paths in reading or saving files. Assume a data subfolder called “Data” and a results subfolder called “Results”. Assuming that the data file resides in the Data subfolder, **your script should run on any computer without errors** and with no need for any change
- Whenever there is a numerical question, (“What is the median of…”), running your code should calculate the required answer, assign it into a variable, and answer the question with a full sentence (i.e. the code should print “The median of something is 17”).
- When asked to remove variables or observations, you should continue the assignment with the reduced data frame.
- Make sure that your file has **meaningful variable names** and is **properly documented** with comments, like in the R scripts presented in class.
- Make sure that all the plots have a proper title, that the x/y axis titled are properly named, or removed if redundant, and that all the labels are clearly shown (i.e. not overlapping)
- Do not forget to answer the questions **for the entire population of Israel** and not just on the sampled population
- Please use only the R libraries presented in class and try to use pipelines as much as possible.

---

1. Load the data file containing answers from the income and expense survey of 2017. Keep only observations from people in the main working age groups.
2. The working age groups contain groups with different number of years. Use the “recode” function to have only 4 age groups, each one with 10 years. Then factorize the variable to make sure that R understands the correct order of the age groups.
3. Show, using geom_bar, a plot of the number of people within each age group. What is the total number of the working age population?
4. Calculate the number of people within each age group, in millions (i.e. 1.1 for 1,100,000). From the resulting data frame, create again the plot for the number of

people within each age group, and add the numbers themselves, in white, at the top of each bar just below the end of the bar (rounded to two significant digits).

5. Create a plot of the share of each population sector with each age group (i.e. a bar per each age group, filled proportionally to the share of each population sector). Separately, calculate analytically the data frame with the 12 values represented in this plot.
6. For the man population only, calculate the employment rate by age group and sector. From this data frame, within a single pipe, create a plot of the employment rate per age group, with each sector in a specific bar, side by side.
   Note: (just for consideration, nothing to submit here) – while the non-Orthodox jews show a typical developed-world pattern of high employment rates up to age 44 with a gradual decline afterwards, the Arab and ultra-Orthodox population shows two different patterns, each reflective of the specific issues for these groups in the labor force.

Analyzing labor income:

7. Filter the data to keep only people with available monthly hours and hourly wage. Based on these two variables, create a new variable for the monthly income, and plot a histogram of the new variable
8. Calculate (within a single pipe), the averages for the monthly income, monthly working hours and hourly wage.
   Think (just for consideration, nothing to submit here): why isn't the average monthly income equal to the multiplication of the average monthly hours and the average hourly wage?
9. Calculate the monthly income by gender and population sector, and create a plot with a facet per sector, and each facet with the average monthly income by gender. Also, add the data as labels on the bars.

Simulation for Orthodox men

We will explore the hypothesis that the low monthly income of Orthodox men is because they work in less lucrative economic branches. We will simulate the change in the average monthly income if Orthodox men worked in economic branches at the same proportion as non-Orthodox Jews, while keeping the same monthly wage per branch. Unfortunately, there are some branches for which there are no Orthodox men in the survey, so we will need to ignore these branches.

10. We will focus on the difference between the monthly income for Orthodox and non-Orthodox Jewish men. Filter the data to keep these two groups only.
11. Create a summary data frame for Orthodox men only, containing (for each economic branch) the average monthly income and the share of people working in this branch. Make sure that the weighted average of the monthly income of all the branches is the same as the average (for Orthodox men) that you received in Q9
12. Look at the help for the function "ntile". Use the function to add a variable to the summary data frame, where for each branch there will be a number identifying the quartile in which its monthly wage appears out of the entire column of monthly wages. Factorize the new variable to give it proper labels and create a plot with the share of employment per branch, where each bar is colored according to the

branch's wage quartile. Note how most employment is concentrated in the two lowest quartiles.

13. Create the same summary data frame as in Q11, for non-Orthodox Jews. However, **before calculating the data frame**, filter and keep only people working in branches that appear in the data frame you created in Q11.
14. (*A bit harder*) Create again the plot from Q12, but add to it, for each branch, on top of the bar, a point marking the share for non-Orthodox Jews. Some hints:
    a. You will need to use one geom_col and one geom_point
    b. As the data for the chart comes from two different data frames, keep the call to ggplot without parameters, and send the data and mapping information explicitly using the geom_ functions

    Advanced – google and find a way to add information in a separate legend that the points mark the non-Orthodox Jews share
15. Create a plot, with the average monthly income of Orthodox Jews per economic branch. Also add a horizontal line marking the average monthly income for the entire population of Orthodox Jews. Assign the plot to a variable before printing it, we will enhance it further in the next questions.
16. Calculate the improved average monthly income for Orthodox man, if their distribution within the economic branches they already work in will be the same as the one for non-Orthodox Jews.
17. Add the line for the improve monthly wage to the plot you created in Q12, with proper color and annotations