Homework 02

DSO 545: Statistical Computing and Data Visualization

Fall 2017

Due Date: Friday September 22, 2017 (11:59 pm)

Instructions

- This homework uses the an air passengers survey data ("college_recent_grads.csv"). You can download it from blackboard.
- Please use R Markdown Documents to answer the following questions.
- Use R, dplyr and ggplot2 to answer all questions. Write R code for each question.
- Make sure to include the R code you used. You won't receive any credit if you don't show the R code chunks.
- Submit your R Markdown and (pdf or word) documentation to blackboard
- I won't tolerate any kind of cheating or late submissions. However I highly encourage to disucss the assignment with each other, but make sure that everyone has a different write up.
- Good luck!

The Guide To Picking A College Major

The millions of American college students heading back to campus this month face a grim reality: A college degree is no guarantee of economic success. But through their choice of major, they can take at least some steps toward boosting their odds. Your task is to help them pick a major by analyzing the following dataset.

The data set `college_recent_grads.csv` contains data about various college majors. **Use `dplyr` and `ggplot2` functions to answer all of the following questions.**

Variables

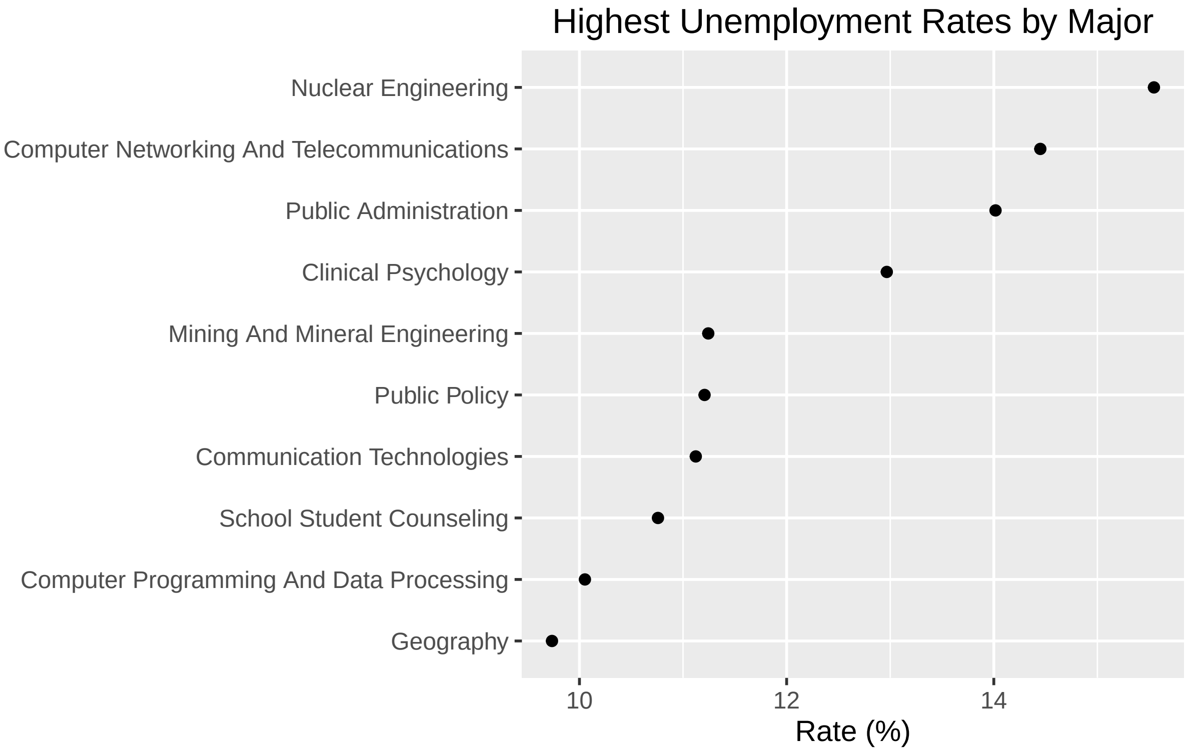
Variable	Description
major	Major description
major_category	Category of major from Carnevale et al
total	Total number of people with major
sharewomen	Percent of the major comprised of women
unemployed	Number unemployed
low_wage_jobs	Number of low wage jobs taken by graduates of the major

(1) Create a dataframe (call it `female_engineering`) for all engineering majors who has 50% or more female students. Sort the dataframe in decreasing order based on the % of female students.

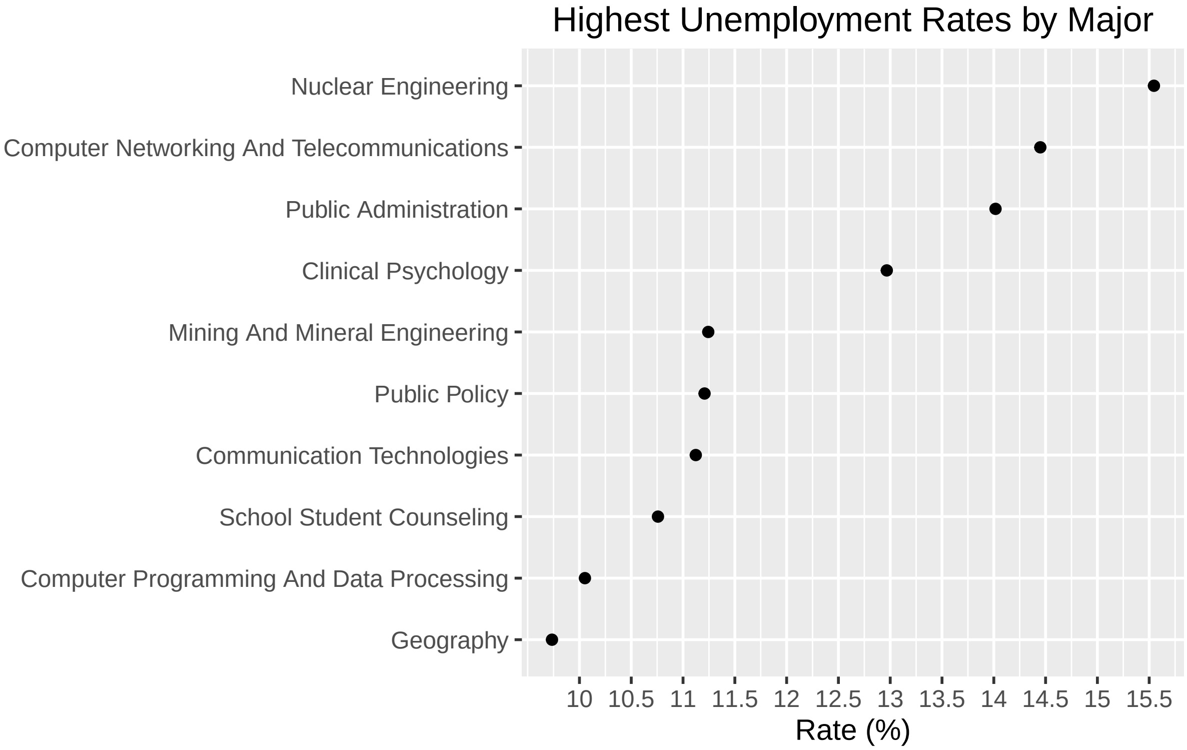
(2) What is the total number (sum) of both engineering and business majors?

(3) Create a new variable `unemployment_rate` (defined as the the number of people unemployed divided by the total), and add it to the original dataset. Then, return a table (call it `top_10`) of the 10 majors with the highest rates as well as those corresponding rates.

(4) Create a dot plot chart to show the highest unemployment rates for the differnt majors.

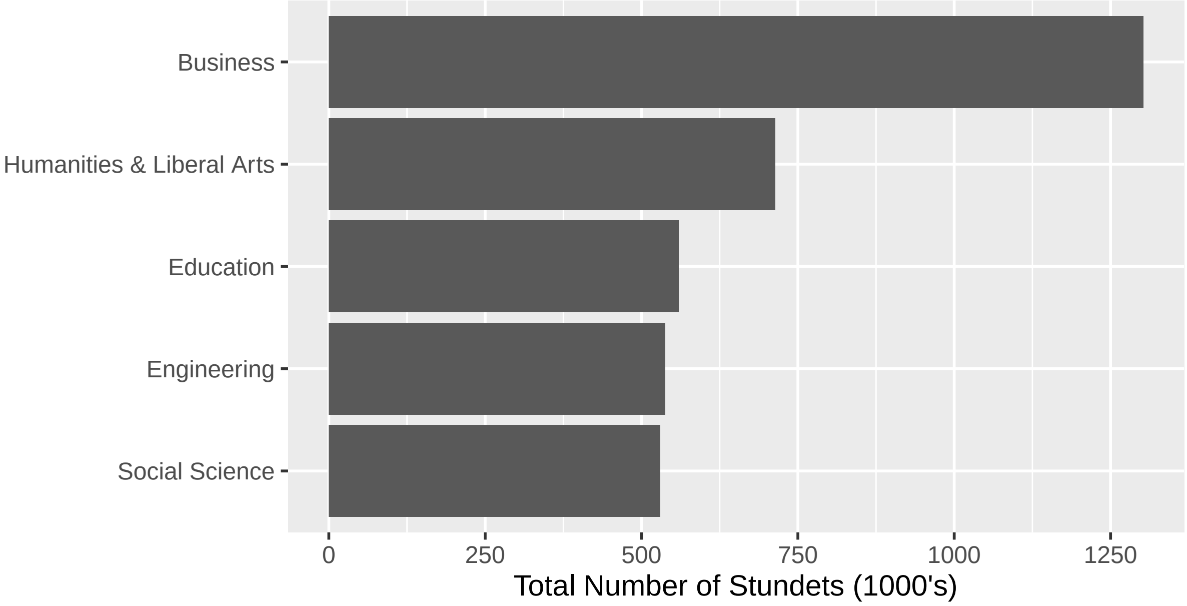


(5) Update the previuos plot as follows (notice the labels on the x-axis). Look online on how to change labels on the x-axis.



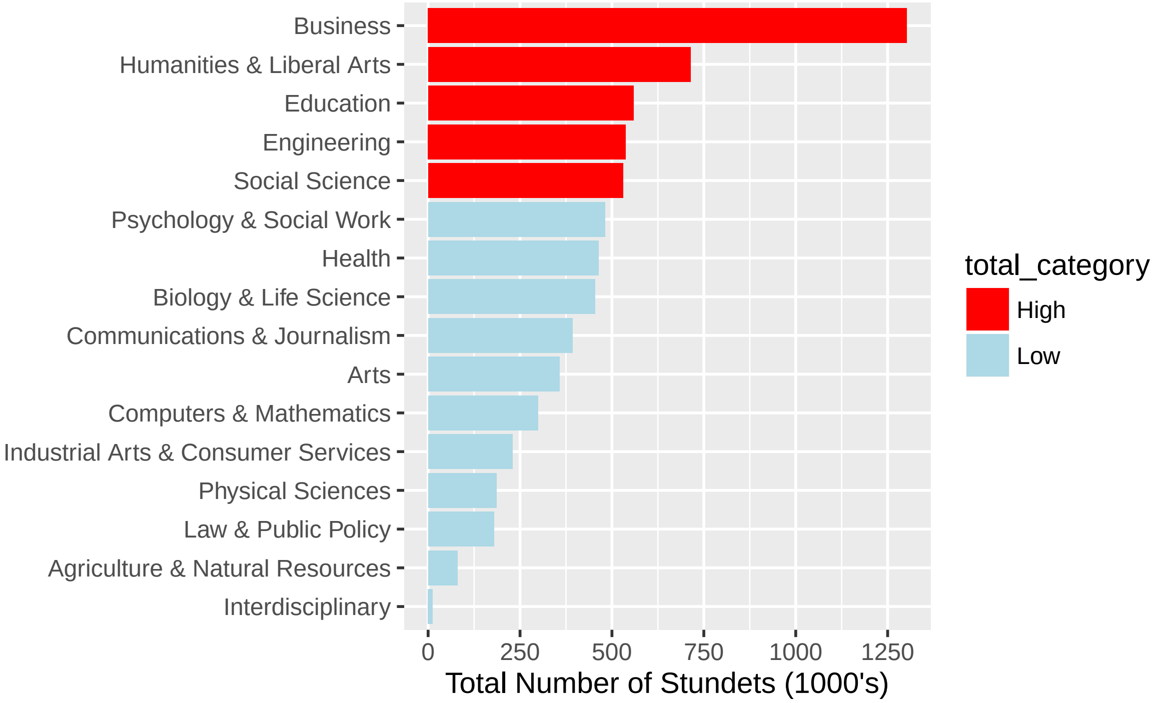
(6) Create a dataframe (call it `majors_total`) that shows the **total** number of students in each of the major categories. Which major category had the highest number of students?

(7) Create an **EXACT** copy of the following graph of the 5 major categories with the most total students.



(8) Using the `majors_total` table you created earlier, create a new variable called `total_category` such that if `total` number of students is less than or equal 500,000, then the category is “Low”, otherwise, it is “High”. (You can use and `if()` or `ifelse()` statement to create the categories based on the specified condition)

(9) Use the `majors_total` to create a copy of the following barchart (High =“red”, Low = “lightblue”):


Business
Humanities & Liberal Arts
Education
Engineering
Social Science
Psychology & Social Work
Health
Biology & Life Science
Communications & Journalism
Arts
Computers & Mathematics
Industrial Arts & Consumer Services
Physical Sciences
Law & Public Policy
Agriculture & Natural Resources
Interdisciplinary
Total Number of Stundets (1000's)
total_category
High
Low