

Homework 3
MTH 3270 Data Science
Due Wed., Feb. 19

Read These Chapters of the Book	Then Do These Exercises
3	Problems 1-2 (below), 3.1, 3.4, 3.8, 3.10*

1 This problem concerns `position` adjustments in "ggplot2".

- a) Look at the help page for `geom_boxplot()`. What's the default `position` adjustment?
- b) Create a visualization of the `mpg` data set (from the "ggplot2" package) using `geom_boxplot()` that demonstrates the default `position` adjustment. Report your R command(s).

2 This problem uses the following data set:

Data Set: nels88

The National Educational Longitudinal Study data set is in the file `nels88.txt`. It is a nationally representative, longitudinal study of 8th graders in 1988 who were followed throughout secondary and postsecondary years.

It included surveys of students reporting on school, work, home experiences, educational resources and support, the role in education of parents and peers, neighborhood characteristics, educational and occupational aspirations, and other student perceptions.

Student assessments were made in reading, social studies, mathematics, and science (8th, 10th, and 12th grades).

The data are from the National Center for Education Statistics,

<https://nces.ed.gov/surveys/nels88/>

It contains 20 variables:

<code>id</code>	Student identifier/student id (unique sample member id)
<code>sch_id</code>	School public release id
<code>heldback</code>	8th grader ever held back a grade
<code>schtype</code>	Eighth grade school type
<code>race</code>	Race
<code>ses</code>	Socio-economic status
<code>female</code>	Sex: female
<code>minority</code>	Student is language minority
<code>asian</code>	Asian/Pacific Islander race
<code>hispanic</code>	Hispanic race
<code>black</code>	Black race
<code>white</code>	White race
<code>native</code>	Native American race
<code>catholic</code>	Catholic religion
<code>private</code>	Private schooled
<code>bymath</code>	Base year (1988) mathematics standardized score
<code>f1math</code>	First follow-up (1990) mathematics standardized score
<code>f2math</code>	Second follow-up (1990) mathematics standardized score
<code>f2dropout</code>	Second follow-up (1990) dropout status

- a) Save the data file `nels88.txt` from the course website, and read it into R using `read.table()` (or `read.csv()`).

Hints: `file.choose()` can be useful for getting the file's location on your computer. Also, specify `header = TRUE` in `read.table()` or `(read.csv())`, and (recommended) `stringsAsFactors = FALSE`.

Explore the data set using `head()`, `str()`, `names()`, `nrow()`, `ncol()`, and `View()`.

How many **rows** does the data set have? How many **columns**? How many of the variables (columns) are **numerical** and how many are **categorical**?

- b) The `nels88` data set has many `NA`'s (missing values). How many of the rows are **"complete"**, i.e. contain **no** `NA`'s?

Hint: Use the function `complete.cases()`.

- c) Fortunately, most of R's graphics functions handle `NA`'s automatically (by eliminating them before plotting the data).

Create at least three different visualizations (graphs) of variables from the `nels88` data set. At least one of your graphs must depict three or more variables.