STAC33

Assignment 1

Due Tuesday January 21 at 11:59pm

My assignments contain two kinds of questions: the ones referring to PASIAS are for your practice, and if you look in PASIAS you will see solutions, which you can check your solution against. The other questions, usually at the end, are an assignment: you need to do these questions yourself and hand them in (instructions below). These are due on the date shown above. An assignment handed in after the deadline is late, and may or may not be accepted (see course outline). I am usually OK with a few minutes late, but not more than that. Don't take that risk.

My solutions to the assignment questions will be available when everyone has handed in their assignment. These are *my* solutions; you can have something different and still be correct.

The grader will have 200-plus assignments to mark, so it is vitally important that you make your assignment *easy* for the grader to deal with. Your answers and explanations *must be easy* for the grader to find. A simple structure is, for each part of each question in order, to put these three things:

- Your *code*
- Your *output*
- Your *answers and explanation*

When you use an R Notebook with your comments below the output, and "preview" the result, the HTML (or Word) file that comes out is in this format. This is why I talk about R Notebooks in class. You can do it differently, but then you have extra work to organize things. *Hand in the HTML file, not the* `.Rmd` *file.* There are no marks for handing in an `.Rmd` file, because *you* need to run your code. You can not expect the grader to run your code on your behalf. That's *your* job. When you hand in your work on Quercus, download it again to check that you handed in what you thought you did.

You are reminded that work handed in with your name on it must be *entirely your own work*. It is as if you have signed your name under it. If it was done wholly or partly by someone else, *you have committed an academic offence*, and you can expect to be asked to explain yourself. The same applies if you allow someone else to copy your work. The graders will be watching out for assignments that look suspiciously similar to each other (or to my solutions). Besides which, if you do not do your own assignments, you *will* do badly on the exams, because the struggle to figure things out for yourself is an important part of the learning process.

Assignments are to be handed in on Quercus. See `https://www.utsc.utoronto.ca/~butler/c32/quercus1.nb.html` for instructions on handing in assignments in Quercus. *Allow yourself time to figure out what you need to do.*

As with any other course involving software, *there are no extensions due to failure to access software.* It is *your* responsibility to make sure that you allow yourself enough time to get connected to R, and to get any packages installed and working. (This is more of an issue if you are using R Studio Cloud, which might be busy and therefore slow; if you install R and R Studio on your own computer, you won't be fighting with other people for resources, but you will need to get everything set up.)

Once you are sitting in front of R Studio (either on `rstudio.cloud` or on your own computer), you would do well to make a new notebook, create a code chunk, in it put

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

and run it (like I just did). If you don't get output like I did, instead getting an error like "no such package", you need to install the Tidyverse first, like this:

```
install.packages("tidyverse")
```

and then you can try again. If at any point R Studio invites you to install any other packages, let it do so.

1. Work through chapter 4 of PASIAS, along with whatever of chapter 3 that you haven't worked through yet.

   Hand the next question in:

2. The Princeton Review conducts an annual survey of high school students who are applying to college, and also of the parents of students who are applying to college. (These are separate surveys; the number of students surveyed is bigger than the number of parents surveyed.) One of the questions was "how far from home would you like the college you attend to be", with students asked to choose one of four categories: less than 250 miles, 250–500 miles, 500–1000 miles, greater than 1000 miles. In the survey that parents completed, the question was "how far from home would you like the college your child attends to be?", with the same distance categories. The data are in `https://www.utsc.utoronto.ca/~butler/assgt_data/distance.csv`. Each row of the data file represents one respondent; the column `who` indicates whether the person responding was one of the students or one of the parents, and the column called `distance` indicates the person's preferred distance category.

   (a) (3 marks) Read in and display (some of) the data. How many people responded to the survey altogether?

   (b) (2 marks) Make a suitable graph of the categorical column `distance` (ignoring `who` for the moment).

   (c) (2 marks) What do you notice that is odd about your bar chart? Why do you think it came out that way? Explain briefly.

   (d) (2 marks) In this case, it makes more sense to arrange the categories in the order they appear in the data. To do this, in place of `distance` use `fct_inorder(distance)` in the code for your graph. Does this put the categories in a sensible order now? (Use `fct_inorder` for the distances in the rest of this question.)

   (e) (2 marks) What does your graph tell you, in the context of the data?

   (f) (2 marks) Make a suitable graph that shows both distances and whether the person involved was a student or a parent. Hints: not stacked, and also you will have to make a decision about which variable is `x` and which is `fill`.

   (g) (3 marks) Another way to make a grouped bar chart is with `position="fill"`. Try that for these data (with `who` as `x`). What has happened? Explain briefly.