Assignment 1

SOSC 13200-2, WIN22

Vincent

Due on 1/17 @ 11:59pm

```
knitr::opts_chunk$set(echo=TRUE)

# load packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

```
# load njmin data
public <- read.table("Datasets/card-krueger_1994/njmin/public.dat",
                     na.strings=c(".","NA"))
```

Overview

You will be working with data used by Nobel-prize winners David Card & Alan Krueger in their famous article on "Minimum wages and unemployment" (1994). You may recall the article from fall quarter. We will re-examine it in Week 2.

Download the "card-krueger_1994" folder and its contents from the "Files/Datasets" section of Canvas. Save the folder to an appropriate location on your computer. (Remember our earlier discussion about directory hygiene.)

The folder contains a folder entitled "njmmin", which contains several files. The relevant files for Assignment 1 are "public.dat" (the dataset) and "codebook.txt" (the codebook).

1 (25 pts)

1.1 In the setup chunk above, load the "public.dat" dataset in to the Environment using the appropriate function for this type of data.

1.2 Write a new code chunk below these lines. In the chunk, use one or two R functions to inspect the data and answer these questions: (a) How many observations (rows) do the data contain? (b) How many columns do they contain? (c) What are the names of the first few variables? (d) Do you think that these variable names are useful? (e) Why or why not?

```
dimensions <- dim(public)
dimensions # this will print out the rows by columns
```

```
## [1] 410  46
```

```
names <- names(public)
names # this will print out all the varialbe names
```

```
##  [1] "V1"  "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11" "V12"
## [13] "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20" "V21" "V22" "V23" "V24"
## [25] "V25" "V26" "V27" "V28" "V29" "V30" "V31" "V32" "V33" "V34" "V35" "V36"
## [37] "V37" "V38" "V39" "V40" "V41" "V42" "V43" "V44" "V45" "V46"
```

1.2 answers:

a. there are 410 rows
b. there are 46 columns
c. the name of the first few variables are V1, V2, V3, V4, V5, and V6
d. These variables are not useful
e. because they are not meaningful and do not tell us anything about what the numbers corresponding to each variable would actually mean

2 (25 pts)

2.1 Examine the codebook for dataset("codebook.txt"). Examine the variable descriptions. Then, create a new code chunk below. In the chunk, rename the variables in the dataset, using the descriptive names given in the codebook.

```
# use the colnames() to rename all the variable names at once:
colnames(public) <- c('SHEET', 'CHAIN', 'CO_OWNED', 'STATE', 'SOUTHJ', 'CENTRALJ',
                      'NORTHJ', 'PA1', 'PA2', 'SHORE', 'NCALLS', 'EMPFT',
                      'EMPPT', 'NMGRS', 'WAGE_ST', 'INCTIME', 'FIRSTINC',
                      'BONUS', 'PCTAFF', 'MEALS', 'OPEN', 'HRSOPEN', 'PSODA',
                      'PFRY', 'PENTREE', 'NREGS', 'NREGS11', 'TYPE2', 'STATUS2',
                      'DATE2', 'NCALLS2', 'EMPFT2', 'EMPPT2', 'NMGR2', 'WAGE_ST2',
                      'INCTIME2', 'FIRSTIN2', 'SPECIAL2', 'MEALS2', 'OPEN2R',
                      'HRSOPEN2', 'PSODA2', 'PFRY2', 'PENTREE2', 'NREGS2', 'NREGS112')
```

NOTE: Upon googling, you will find that there are various functions for renaming variables in R. Some are more efficient than others. You may use whichever you prefer.

I realized there is a way to read from text file and extract all the names, but because there are only 46 variables, its probably less energy to just type them all up, but I might revisit some time to recode it! :D

3 (25 pts)

3.1 View the dataset using the `View()` function or by clicking on the dataset in the Environment tab. What character seems to denote missing data?

```
View(public)
```

looking at the data, I notice that period (".") represents missing data

3.2 Click on the small blue circle with white arrow in the Environment tab. Note that some numeric variables were imported as character variables upon loading the dataset. Return to where you loaded the dataset in the setup chunk. Add some options to your loading function to (i) import missing data as `NA` and (ii) import variables as their proper type.

4 (25 pts)

4.1 Here is Card & Krueger's (1994) description of their Full-time-equivalent (fte) employment variable (p. 775):

"Full-time-equivalent (FTE) employment was calculated as the number of full-time workers [including managers] plus 0.5 times the number of part-time works."

Create a new code chunk below. In the chunk, replicate the authors' FTE variable. Create the variable as new variable *inside* the dataset.

```
# 12 and 13 and the column numbers of EMPFT and EMPPT respectively
public$FTE <- (public[,12] + (0.5*public[,13]))
```

4.2 Check you work by printing the first few values of your new variable in the code chunk below.

```
# These are the FTE variables
head(public[,47], 10)
```

```
##  [1] 37.50  9.75  6.50 30.00 19.00 15.50 67.50 18.50  6.00  7.00
```

```
# Let us compare them to the EMPFT and EMPPT to see if the math works out
head(public[,12], 10)
```

```
##  [1] 30.0  6.5  3.0 20.0  6.0  0.0 50.0 10.0  2.0  2.0
```

```
head(public[,13], 10)
```

```
##  [1] 15.0  6.5  7.0 20.0 26.0 31.0 35.0 17.0  8.0 10.0
```