Problem set 1 question 1

Juntong Lin

2020-10-02

Question 1

part(a)

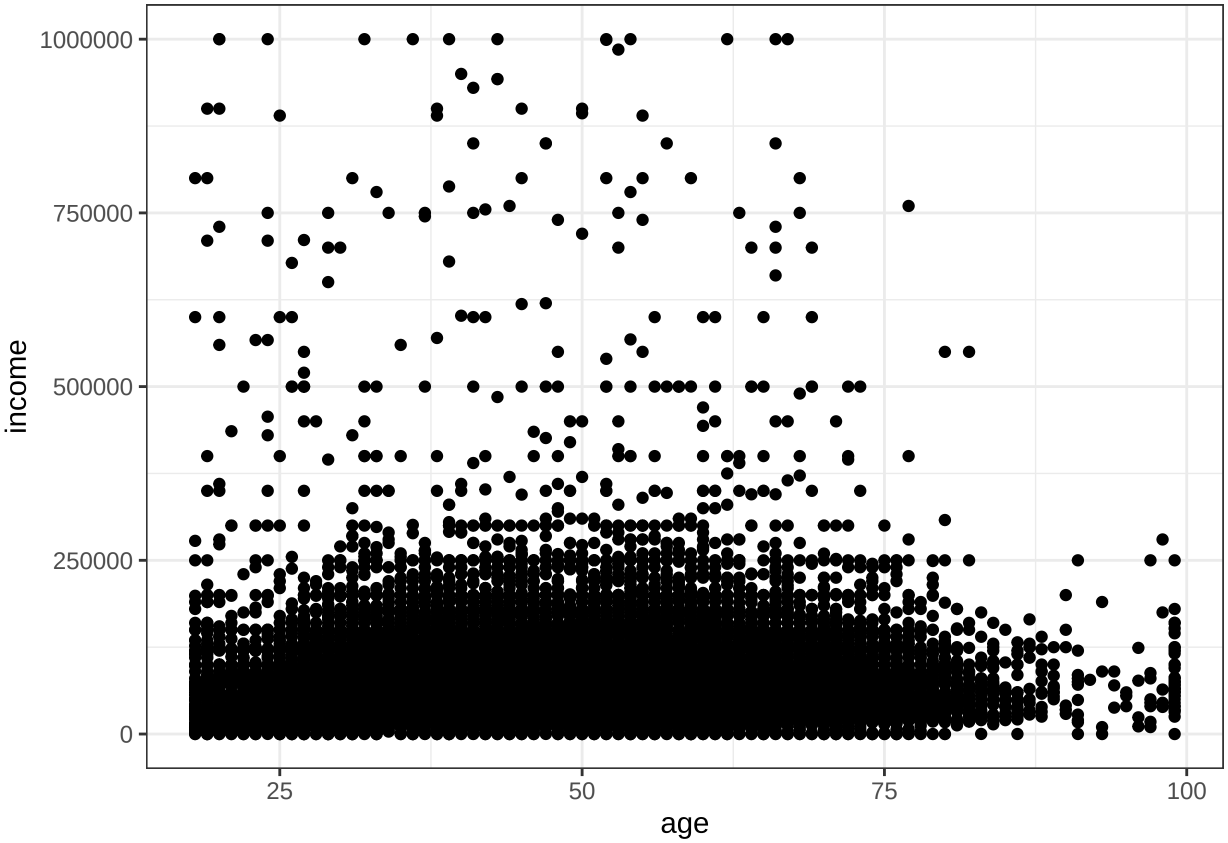
```
install.packages("opendatatoronto")
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
install.packages("devtools")
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
devtools::install_github("sharlagelfand/opendatatoronto")
## Downloading GitHub repo sharlagelfand/opendatatoronto@HEAD
##
##      checking for file '/tmp/Rtmpktu9jy/remotes17733824b6be/sharlagelfand-opendatatoronto-0f65775/DE
##   -  preparing 'opendatatoronto':
##      checking DESCRIPTION meta-information ...  v  checking DESCRIPTION meta-information
##   -  checking for LF line-endings in source and make files and shell scripts
##   -  checking for empty or unneeded directories
##   -  building 'opendatatoronto_0.1.3.9001.tar.gz'
##
##
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
devtools::install_github("hodgettsp/cesR")
## Skipping install of 'cesR' from a github remote, the SHA1 (7c780beb) has not changed since last inst
##   Use `force = TRUE` to force installation

library(opendatatoronto)
library(cesR)
library(tidyverse)
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
library(visdat)
library(skimr)
```

```
cesR::get_decon()
## TO CITE THIS SURVEY FILE: Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John
##              https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1
## LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V
```

The data is a research of 22 variables, such as citizenship, age and income, for the people on ces2019_web. There are total 37822 observations and 22 variables. The dataset is interesting because the population is large, and various features are researched. Thus I can study the relationship between many different kinds of fields.

part(b)

```
decon <- decon %>%  mutate(Age = 100 - as.numeric(decon$yob))
decon %>% filter(income <= 1000000) %>%
  ggplot(aes(x =Age, y = income)) + geom_point() + theme_bw() + labs(x ="age", y = "income")
```



I made a scatter plot for age and income, with age on the x axis and income on the y axis. From the plot, most of the observations gather under the 250000. This means no matter the age, most of the people have income under 250000. Overall, the plot shows a weak positive linear trend. The age has a weak influence on the income, and as getting older, the average income would increase a little bit.

part(c)

```
decon %>% group_by(income_cat) %>% summarize(mean_age = mean(Age), median_age = median(Age))
## `summarise()` ungrouping output (override with `.groups` argument)
## # A tibble: 10 x 3
##    income_cat                       mean_age median_age
##    <fct>                               <dbl>      <dbl>
##  1 No income                            31.0         26
##  2 $1 to $30,000                        44.3         44
##  3 $30,001 to $60,000                   47.9         48
##  4 $60,001 to $90,000                   47.0         46
##  5 $90,001 to $110,000                  45.8         45
##  6 $110,001 to $150,000                 45.2         45
##  7 $150,001 to $200,000                 46.0         46
##  8 More than $200,000                   46.3         49
##  9 Don't know/ Prefer not to answer     49.2         51
## 10 <NA>                                 49.6         50
```

I found that income generally increase as age increase. Though the age for groups with income more than 30000 fluctuate, the no income group and 1 to 30,000 group is obviously younger. And the highest group has the oldest mean age.

part(d)

In this dataset, I was interested in the relationship between the age and income. I estimate that the income will increase as getting older before getting the data. After study the scatter plot, I find that the positive linear relationship is weak, because most of the people at all age range gather under a value. And from the summarize study, the mean age in the middle groups fluctuate a lot. Thus the relationship between this two variables is actually complicated.

part(e)

[1] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[2] Tierney N (2017). “visdat: Visualising Whole Data Frames.” *JOSS*, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.

[3] Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible Summaries of Data. https://docs.ropensci.org/skimr (website), https://github.com/ropensci/skimr.