

Lecture 3 Problem Set

INSERT YOUR NAME HERE

INSERT DATE HERE

In this problem set, you will investigate objects and data patterns via tidyverse

Step 1: Investigate Missing Observations

Load tidyverse package [code already provided]

```
#install.packages("tidyverse") #install if you do not have tidyverse installed
library(tidyverse)
#> -- Attaching packages -----
#> v ggplot2 3.3.2      v purrr 0.3.4
#> v tibble 3.0.3       v dplyr 1.0.2
#> v tidyr 1.1.1        v stringr 1.4.0
#> v readr 1.3.1        v forcats 0.5.0
#> -- Conflicts -----
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag()     masks stats::lag()
```

Load data frame object and use count() to tabulate the total number of visits each school received [code already provided]

```
rm(list = ls()) # remove all objects
getwd()
#> [1] "/Users/karinasalazar/rclass/lectures/lecture3"

load(url("https://github.com/ozanj/rclass/raw/master/data/recruiting/recruit_school_allvars.RData"))
#glimpse(df_school_all)

df_school_all %>% count(total_visits)
#> # A tibble: 25 x 2
#>   total_visits     n
#>   <int> <int>
#> 1         0 15405
#> 2         1  2718
#> 3         2  1324
#> 4         3   671
#> 5         4   395
#> 6         5   263
#> 7         6   152
#> 8         7   107
#> 9         8    89
#> 10        9    57
#> # ... with 15 more rows
```

1. In these questions, apply the `table()` function to the variable `school_type` within the object `df_school_all`. Run the appropriate R code within the chunk below and write substantive responses below each question.
 - In your own words, what does the `table()` function do?
 - What does the `useNA` argument of the `table()` function control?
 - What is the default value of the `useNA` argument and what does this default value mean?
 - What happens when you assign the value "ifany" to the `useNA` argument?
 - What happens when you assign the value "always" to the `useNA` argument?
 - In the below R code chunk, use the `table()` function to count the number of observations for each value `school_type` three different ways:
 - First, without specifying any value for `useNA`
 - Second, by assigning the value "ifany" to the `useNA` argument
 - Third, by assigning the value "always" to the `useNA` argument

Step 2: Filter, select, arrange questions

The data frame `df_school_all` has one observation for each high school (public and private).

- The variables that begin with `visits_by_...` identify how many off-campus recruiting visits the high school received from a particular public university. For example, UC Berkeley has the ID 110635 so the variable `visits_by_110635` identifies how many visits the high school received from UC Berkeley.
- The variable `total_visits` identifies the number of visits the high school received from all (16) public research universities in this data collection sample.

For the questions below, imagine that you have been asked by a major news outlet to identify which high schools receive the most total number of off-campus recruiting visits from public universities.

- For all questions below, you can answer using one line of code or you can answer in several steps (e.g., first create new data frame, then print selected variables)
 - For questions that ask you to print the “top 10” observations, you can simply print the object and rely on the fact that the default option [for “tibble” data frames] is to print 10 observations OR you can wrap the command in the `head()` function and explicitly tell R to print 10 observations
 - Before conducting analyses, we’ll rename the variable `avgmedian_inc_2564` to give it a shorter name.
1. Rename the variable `avgmedian_inc_2564` to `med_inc` and assign new variable name to the existing object `df_school_all`
 2. The news outlet is interested in comparing the in-state and out-of-state high school visits for The University of Alabama (IPEDS ID = 100751) variable `visits_by_100751`. Count the number of in-state public high schools that received at least one visit from The University of Alabama.
 - Note: You will need to use `filter` and the `count` function
 - Use commas to separate variables for this question (e.g. `filter(dataframe, variable == something, variable == something)`)
 - You can do this in one step by wrapping the `count` function around the `filter` function ; or you can do this in two steps by creating a new data frame first
 3. How many public in-state high schools visited by The University of Alabama enroll at least 50% Latinx students **or** 50% Black students?
 - hint: use the variables `pct_hispanic` and `pct_black`.

4. Now count the out-of-state public high schools that received at least one visit by the University of Alabama without using commas to separate conditions in your filter.

- hint: `&` or `%in%`

5. How many public out-of-state high schools visited by The University of Alabama enroll at least 50% Latinx students **or** 50% Black students?

Once finished, knit to (HTML) and upload both .Rmd and HTML files to ps_submit_2 on Piazza

Remember to use this naming convention "lastname_firstname_ps2"