Lecture 5 problem set

INSERT YOUR NAME HERE

October 25, 2019

Contents

Purpose

Data you will be working with

In this problem set, we are working with data from the the list of prospective students that Western Washington University purchased from College Board. We have also merged in Census data on socioeconomic/racial characteristics and NCES data on school characteristics to the prospect-level data from College Board. Hence, the dataset you will be working with has one observation per prospect (i.e., student). Some variables are prospet-level variables (e.g., `ethn_code` is a measure of race/ethnicity that varies by prospect). Other variables measured at the zip-code level or state-level. These are measures of the racial composition for the zip code the prospect lives in and measures of the racial composition for the state in which the prospect lives; they do not vary across prospects within the same zip-code or state.

Task

For this problem set, you are a researcher and your goal is to identify systematic racial and socioeconomic bias in student list purchases by Western Washington University. That is, do the prospects purchased by Western Washington tend to have different racial and socioeconomic characteristics than other people in their state or zip-code?

Note that there is a lot of data cleaning required before conducting `group_by` and `summarise()` analyses. Much of this data cleaning involves creating prospect-level and zipcode/state-level measures of race/ethnicity that are consistent to one another. Therefore, we have answered some of the data cleaning questions for you to avoid making the problem set too long. We intentionally left our data cleaning code for you all to get a sense of the process of investigating and cleaning your data.

Note, for questions that ask you to use `summarize()` function, fine to use `summarize_all()`, `summarize_at()`, or `summarize_if()` instead as long as you get the right answer.

Caveat

Merging data from other sources (e.g. College Board & Census) gives us breadth in investigating racial and socioeconomic bias beyond the prospect (student) level, yet at the same time, we are limited in the choices we make for disaggregating by race and ethnicity (in addition to other variables). Further, there are some fundamental differences between how College Board and Census define race/ethnicity that cannot be overcome with data cleaning. Therefore, comparisons between race/ethncity variables from College Board and race/ethnicity variables from Census are problematic.

Definitions for race and ethnicity used by Census and College Board

Here is some background information on how U.S. Census and College Board define race and etncity:

- U.S. Census
  - Census efinitions of race and ethnicity LINK HERE
  - Census categories of race and ethnicity LINK HERE
- College Board
  - College Board Categories of race and ethnicity LINK HERE
  - College Board race and ethnicity questions from SAT Questionnaire LINK HERE

Idiosyncracies about the way race/ethnicity is defined by College Board vs. U.S. Census in the dataset you will be working with

- The College Board survey asks a question about "ethnicity" and then a separate question about "race"; However, the data sent to us by Western Washington combined race and ethnicity into one variable called `ethn_code`
- The College Board survey questions for ethnicity and race uses the following rules:
  - "Students may select all options that apply. In prior years, they were asked to select one option."
- By contrast, US Census data asks respondents to select one option; there is a separate option for "Two or More Races"

- As a result of these differences, the College Board race/ethnicity variable has a much higher percentage of people who identify as "2 or more races" than data from U.S. Census

Load library and data

```
library(tidyverse)
#> -- Attaching packages ---------------------------------------------------------------------------- t
#> v ggplot2 3.2.1     v purrr   0.3.2
#> 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 ------------------------------------------------------------------------------- tidyver
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag()    masks stats::lag()
```

```
rm(list = ls()) # remove all objects

load(url("https://github.com/ozanj/rclass/raw/master/data/prospect_list/wwlist_merged.RData"))
#getwd()
#load("../../../documents/rclass/data/prospect_list/wwlist_merged.RData")
```

Cleaning the data before creating summary measures using group_by() and summarise()

In general, for all questions that ask you to drop certain observations or create new variables, assign these changes to the existing object `wwlist`

Part I: Questions related to keeping/dropping specfic observations

Question 1

- Do the following:
    - Count the number of observations that have `NA` for the variable `state`
    - Using `filter()` drop all observations that have `NA` for the variable `state`
    - Using `mutate()` and `if_else()`, create a [and retain] 0/1 variable `in_state` that equals 1 if `state` equals Washington and equals `0` otherwise
    - Investigate the values of the new variable `in_state`, including confirming that this variable has no missing values

Question 2

- Do the following:
    - Count the number of observations where the value of `pop_total_zip` equals `0`
    - Count the number of observations where the value of `pop_total_zip` equals `NA`
    - Drop observations where the value of `pop_total_zip` is equal to `0`
        - ∗ NOTE: we won't drop observations where value of `pop_total_zip` equals `NA`

NOTE: IN THIS QUESTION, WE GIVE YOU THE ANSWERS; ALL YOU HAVE TO DO IS RUN THE BELOW CODE CHUNK

```
wwlist %>% filter(pop_total_zip ==0) %>% count() # number of obs that equal 0
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1    23
wwlist %>% filter(is.na(pop_total_zip)) %>% count() # number of obs that equal NA
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1641

wwlist %>% filter(pop_total_zip != 0 | is.na(pop_total_zip)) %>%
  count() # number of obs where pop_total zip is either not equal to 0 or is equal to NA
#> # A tibble: 1 x 1
#>        n
#>    <int>
#> 1 268373

wwlist <- wwlist %>%
  filter(pop_total_zip != 0 | is.na(pop_total_zip)) # keep obs where pop_total_zip is not equal to 0 or
```

Question 3

- Remove observations the have the following values for the variable `state`: “AP”, “MP”
    - these values either refer to territories or are errors

NOTE: IN THIS QUESTION, WE GIVE YOU THE ANSWERS; ALL YOU HAVE TO DO IS RUN THE BELOW CODE CHUNK

```
wwlist %>% filter(state %in% c("AP","MP")) %>% count() # equal to AP or MP
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1     2
wwlist %>% filter(!state %in% c("AP","MP")) %>% count() # not equal to AP or MP
#> # A tibble: 1 x 1
#>        n
#>    <int>
#> 1 268371

wwlist <- wwlist %>% filter(!state %in% c("AP","MP")) # not equal to AP or MP
wwlist %>% count(state)
#> # A tibble: 52 x 2
#>    state     n
#>    <chr> <int>
#>  1 AK     3671
#>  2 AL      136
#>  3 AR       78
#>  4 AZ    10358
#>  5 CA    62382
#>  6 CO    24822
#>  7 CT      173
#>  8 DC       35
#>  9 DE       37
#> 10 FL     1287
#> # ... with 42 more rows
```

Part II: Questions related to creating new variables prior to creating summary measures using group_by() and summarise()

This set of questions primarily relates to creating prospect-level measures of race/ethnicity (data from College Board) that are consistent with zip-code-level and state-level measures of race/ethnicity (data from US Census)

Question 1

- Investigate the prospect-level race/ethnicity variable `ethn_code` as follows:
  - what "type" of variable is it
  - create a frequency table
  - count the number of `NA` values

NOTE: IN THIS QUESTION, WE GIVE YOU THE ANSWERS; ALL YOU HAVE TO DO IS RUN THE BELOW CODE CHUNK

```
str(wwlist$ethn_code)
#>  chr [1:268371] "other-2 or more" "white" "white" "other-2 or more" ...
wwlist %>% count(ethn_code)
#> # A tibble: 10 x 2
#>    ethn_code                                               n
#>    <chr>                                               <int>
#>  1 american indian or alaska native                      202
#>  2 asian or native hawaiian or other pacific islander   2385
#>  3 black or african american                             563
#>  4 cuban                                                  70
#>  5 mexican/mexican american                             6549
#>  6 not reported                                         5737
#>  7 other spanish/hispanic                               2431
#>  8 other-2 or more                                     90579
#>  9 puerto rican                                          195
#> 10 white                                              159660
wwlist %>% filter(is.na(ethn_code)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1     0
```

Question 2

- The prospect-level variable `ethn_code` combines Asian, Native Hawaiian and Pacific Islander into one category. To be consistent with the prospect-level variable `ethn_code`, create a variable `pop_api_zip` equal to the sum of `pop_asian_zip` and `pop_nativehawaii_zip`. Follow these steps:
  - check how many missing values the "input variables" `pop_asian_zip` and `pop_nativehawaii_zip` have
  - create the new variable
  - check the value of the new variable for observations that had missing values in the input variables
  - delete the input variables

NOTE: IN THIS QUESTION, WE GIVE YOU THE ANSWERS; ALL YOU HAVE TO DO IS RUN THE BELOW CODE CHUNK

```
#investigate input variables [zip-code level race/ethnicity vars]
wwlist %>% filter(is.na(pop_asian_zip)) %>% count()
```

```
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_nativehawaii_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639

#create variable
wwlist <- wwlist %>% mutate(
    pop_api_zip = pop_asian_zip + pop_nativehawaii_zip
  )

#check value of new variable; and check the value of the new variable against value of input variables
wwlist %>% filter(is.na(pop_api_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_asian_zip)) %>% count(pop_api_zip)
#> # A tibble: 1 x 2
#>   pop_api_zip     n
#>         <int> <int>
#> 1          NA  1639
wwlist %>% filter(is.na(pop_nativehawaii_zip)) %>% count(pop_api_zip)
#> # A tibble: 1 x 2
#>   pop_api_zip     n
#>         <int> <int>
#> 1          NA  1639

#remove input variables
wwlist <- wwlist %>% select(-pop_asian_zip,-pop_nativehawaii_zip)

#names(wwlist)
```

Question 3

- Follow the same steps as above to create a variable `pop_api_state` from the input variables

Question 4

- Next, we'll use the zip-code level measures of number of people by race/ethnicity to create zip-code level measures of **percent** of people by race/ethnicity
  - Before creating the new variables, investigate presence of missing observations in input variables
  - after you create the variables, investigate the value of the new variables and their value against missing values of the input variables. Do this for two of the new race variables you created

NOTE: IN THIS QUESTION, WE GIVE YOU THE ANSWERS; ALL YOU HAVE TO DO IS RUN THE BELOW CODE CHUNK

```
#show names of zip code level race vars
wwlist %>% select(ends_with("_zip"),-med_inc_zip) %>% names()
#> [1] "pop_total_zip"      "pop_white_zip"      "pop_black_zip"
#> [4] "pop_latinx_zip"    "pop_nativeam_zip"  "pop_multirace_zip"
#> [7] "pop_otherrace_zip" "pop_api_zip"

#Investigate presence of missing values in input variables
wwlist %>% filter(is.na(pop_total_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_white_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_black_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_latinx_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_nativeam_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_multirace_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_otherrace_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_api_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639

#create new variables
  #note: we multiply by 100 so that we have percentages rather than proportions, which are easier to re
wwlist <- wwlist %>%
  mutate(
    pct_white_zip= pop_white_zip/pop_total_zip*100,
    pct_black_zip= pop_black_zip/pop_total_zip*100,
    pct_latinx_zip= pop_latinx_zip/pop_total_zip*100,
    pct_nativeam_zip= pop_nativeam_zip/pop_total_zip*100,
    pct_multirace_zip= pop_multirace_zip/pop_total_zip*100,
    pct_otherrace_zip= pop_otherrace_zip/pop_total_zip*100,
    pct_api_zip= pop_api_zip/pop_total_zip*100,
  )

#Investigate values of new variables against values of input vars for two of the race categories

wwlist %>% summarise(pct_white_zip= mean(pct_white_zip, na.rm = TRUE)) # average percent white across a
#> # A tibble: 1 x 1
#>   pct_white_zip
#>           <dbl>
#> 1          68.0

wwlist %>% filter(is.na(pct_white_zip)) %>% count() # number missing
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_white_zip) | is.na(pop_total_zip)) %>%
  count(pct_white_zip) # count values of pct_white_zip if either of the input vars is missing
#> # A tibble: 1 x 2
#>   pct_white_zip     n
#>           <dbl> <int>
#> 1            NA  1639

wwlist %>% filter(is.na(pct_black_zip)) %>% count()
#> # A tibble: 1 x 1
#>       n
#>   <int>
#> 1  1639
wwlist %>% filter(is.na(pop_black_zip) | is.na(pop_total_zip)) %>%
  count(pct_white_zip)
#> # A tibble: 1 x 2
#>   pct_white_zip     n
#>           <dbl> <int>
#> 1            NA  1639
```

Question 5

- Follow the same steps as above to create state-level measures of percent of people by race/ethnicity
    - after you create the variables, investigate the value of the new variables and their value against missing values of the input variables for two of the new race variables

Question 6

- Next, we'll make a new version of the prospect level race/ethnicity variable that is consistent with the Census zip code level and state level race/ethnicity variables
    - First, investigate the input variable `ethn_code` including:
        - identifying variable "type"
        - creating a frequency table
        - counting the number of missing values
    - Second, Using the `recode()` function within `mutate()`, create a variable called `ethn_race` that recodes the input variable `ethn_code` as follows:
        - "american indian or alaska native" = "nativeam",

        - “asian or native hawaiian or other pacific islander” = “api”,
        - “black or african american” = “black”,
        - “cuban” = “latinx”,
        - “mexican/mexican american” = “latinx”,
        - “not reported” = “not_reported”,
        - “other-2 or more” = “multirace”,
        - “other spanish/hispanic” = “latinx”,
        - “puerto rican” = “latinx”,
        - “white” = “white”,

    - Third, investigate the values of the new variable `ethn_race` including:
        - variable type
        - creating a frequency table
        - counting the number of missing values
        - Then run this code to check the values of the new variable against the values of the input variable:
        - `wwlist %>% group_by(ethn_race) %>% count(ethn_code)`

Question 7

- Based on the variable `ethn_race` you just created, create a set of 0/1 prospect-level race indicator indicators
- `nativeam_stu`; `api_stu`; `black_stu`; `latinx_stu`; `multirace_stu`; `white_stu`, `notreported_stu`
- after creating the 0/1 indicators check their values against the value of the input variable

NOTE: IN THE BELOW CODE CHUNK, I’LL CREATE THE INDICATOR FOR `nativeam_stu`; YOU CREATE THE REMAINING
Uncomment this code chunk after creating the `ethn_code` variable from the code chunk above

```
#wwlist %>% count(ethn_race)
#wwlist %>% count(ethn_code)

#Create var
#wwlist <- wwlist %>%
#  mutate(nativeam_stu = ifelse(ethn_race == "nativeam",1,0))

#Investigate var
#wwlist %>% count(nativeam_stu)
#wwlist %>% group_by(nativeam_stu) %>% count(ethn_race)
```

Part III: group_by() and summarise() questions

Now that we have cleaned data and created variables in prospect-level dataset, we can use `group_by()` and `summarise()` to perform calculations across rows about the characteristics of prospects purchased and how they compare to the general population. Generally, for the below questions you don’t need to retain/assign the object created by `group_by()` and `summarise()`

Question 1

- Grouping by the variable `in_state`, use `summarise()` to create the following measures:
    - `tot_prosp`: a count of the number of prospects purchased

Question 2

- Grouping by the variable `in_state`, use `summarise()` to create the following measures:

  - `tot_prosp`: a count of the number of prospects purchased
  - `white`: a count of number of white prospects purchased, based on the input var `white_stu`
    - **hint:** `newvar = sum(input_var, na.rm=TRUE)`

Question 3

- Grouping by the variable `in_state`, use `summarise()` to create the following measures:
  - `tot_prosp`: a count of the number of prospects purchased
  - `report_race`: the total number of prospects purchased that reported race (**hint:** `sum(ethn_race !="not_reported", na.rm=TRUE)`)
  - `white`: a count of number of white prospects purchased, based on the input var `white_stu`

Question 4

- Grouping by the variable `in_state`, use `summarise()` to create the following measures:
  - `tot_prosp`: a count of the number of prospects purchased
  - ‘report_race: the total number of prospects purchased that reported race

  - a count of number of prospects purchased by race based on each of the following input variables (that is, you will create 7 variables)
    - **nativeam_stu , api_stu , black_stu , latinx_stu , multirace_stu , white_stu , notreported_stu**

Question 5

- Grouping by the variable `in_state`, use `summarise()` to create the following measures:
- `tot_prosp`: a count of the number of prospects purchased
- `white`: a count of number of white prospects purchased, based on the input var `white_stu`

- `p_white`: the proportion of prospects purchased that were white for each by group, based on the 0/1 input var `white_stu`

- **hint:** `newvar = mean(input_var, na.rm=TRUE)`

Question 6

- Grouping by the variable `in_state`, use `summarise()` to create the following measures:
- `tot_prosp`: a count of the number of prospects purchased
- the **percent** of prospects purchased from each race group based on the following 0/1 indicator variables (that is, you will create 7 variables)
  - **nativeam_stu , api_stu , black_stu , latinx_stu , multirace_stu , white_stu , notreported_stu**
  - **hint:** since you are creating **percent** measures rather than **proportion**: `newvar = mean(input_var)*100`

Question 7

- Now we will group_by the variable `state` (rather than `in_state`), use `summarise()` to create the following measures:
  - `tot_prosp`: a count of the number of prospects purchased
  - `white`: a count of number of white prospects purchased, based on the input var `white_stu`

  - `p_white`: the **percent** of prospects purchased that were white for each by group, based on the 0/1 input var `white_stu`

Part IV: Comparing prospects purchased to regional income and racial demographics

Question 1

In this question, we will compare median zip code income of prospects purchased to the median income in the states they live in. The goal is to assess whether Western Washington is disproportionately purchasing more affluent prospects. The variable `med_inc_state` identifies the median income of all people in the state aged 25-64. This variable has the same value for all prospects in the same state. Therefore, when using `group_by()` and `summarise()`, we can just grab the first observation for each state (hint: `first(input_var)` or `nth(input_var,1)`).

To answer this question, group_by **`state`** and use `summarise()` to create the following measures:

- `tot_prosp`: a count of the number of prospects purchased

- `med_inc_zip_stu`: the mean value of the variable `med_inc_zip` for each by group

- `med_inc_state`: the first value of the variable `med_inc_state` for each by group

Question 2

For each state, we want to compare the percent of prospects purchased who are white to the percent of people in the state who are white. The variable `pct_white_state` identifies the percent of people in the state who are white. This variable has the same value for all prospects in the same state. Therefore, when using `group_by()` and `summarise()`, we can grab the first observation for each state (hint: `first(input_var)` or `nth(input_var,1)`).

- group_by **`state`** and use `summarise()` to create the following measures:
  - `tot_prosp`: a count of the number of prospects purchased
  - `white`: a count of number of white prospects purchased, based on the input var `white_stu`

  - `p_white`: the **percent** of prospects purchased that were white for each by group, based on the 0/1 input var `white_stu`
  - `p_white_st`: the percent of people in the state who are White, based on the input variable `pct_white_state`

Question 3

- group_by **`state`** and use `summarise()` to create the following measures:
  - `tot_prosp`: a count of the number of prospects purchased
  - Create (A) a measure of the percent of prospects who identify as a particular race/ethnicity group and (B) the percent of people in the state who identify as that particular race/ethnicity group for the following race/ethnicity groups: **multirace, white, api, black, latinx**

Question 4

- The goal of this question is to compare the race of prospects purchased from Washington to the racial composition of zip-codes in Washington. For this question, you will filter to **only include prospects who are from Washington AND do not have the value `NA` for the variable `pop_total_zip`**, then group by the variable `zip5` and use `summarise()` to create the following variables:
  - `tot_prosp`: a count of the number of prospects purchased
  - Create (A) a measure of the percent of prospects in the zip-code who identify as a particular race/ethnicity group and (B) the percent of people in the zip-code who identify as that particular race/ethnicity group for the following race/ethnicity groups: **multirace, white, api, black, latinx**

Once finished, knit to (pdf) and upload both .Rmd and pdf files to class website under the week 4 tab
Remeber to use this naming convention "lastname_firstname_ps4"