Brilliant Cassowary

Exploratory data analysis

Nidhi Soma (ns848) Joice Chen (jc3528) Jinpeng Li (jl3496)
Stephen Syl-Akinwale (sis33)

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(janitor)
```

```
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```
library(dplyr)
library(skimr)
```

Inserting the dataset on Coffee

```
coffee_df<-read_csv("data/GACTT_RESULTS_ANONYMIZED_v2.csv")
```

```
Rows: 4042 Columns: 113
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (44): Submission ID, What is your age?, How many cups of coffee do you t...
dbl (13): Lastly, how would you rate your own coffee expertise?, Coffee A - ...
lgl (56): Where do you typically drink coffee? (At home), Where do you typic...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Research question(s)

Research question(s). State your research question (s) clearly.

Data collection and cleaning

Have an initial draft of your data cleaning appendix. Document every step that takes your raw data file(s) and turns it into the analysis-ready data set that you would submit with your final project. Include text narrative describing your data collection (downloading, scraping, surveys, etc) and any additional data curation/cleaning (merging data frames, filtering, transformations of variables, etc). Include code for data curation/cleaning, but not collection.

1. After inserting the data, it is in a table format. We need to evaluate which columns are useful or pretty much empty. This section removes columns with mostly NA values, since they will not be helpful for analysis.

```
#remove NA columns
coffee_clean <- coffee_df |>
  select(-contains("flavorings")) |>
  select(-contains("Gender (please specify"))

#new names
#coffee_clean <- coffee_df |>
#  rename_with(~str_extract(.x, '(?<=\\().*?(?=\\))'))
```

2. Next, we need to evaluate which columns contain repetitive information. There are a few columns that ask a question, and the next few columns are the answer to that question, so the question itself contains repetive values that we already have in the subsequent column names. This section removes the question columns.

```
#remove repetitive questions
coffee_clean <- coffee_clean |>
  mutate(`Where do you typically drink coffee?` = NULL) |>
  mutate(`How do you brew coffee at home?` = NULL)|>
  mutate(`On the go, where do you typically purchase coffee?` = NULL) |>
  mutate(`Do you usually add anything to your coffee?` = NULL) |>
  mutate(`What kind of diary do you add?` = NULL) |>
  mutate(`What kind of sugar or sweetener do you add?` = NULL) |>
  mutate(`Why do you drink coffee?` = NULL)
```

3. The main part of our data cleaning is fixing the column names to be in a tidy format. We go through and rename columns in the original form of "question? (response)" to "question_response". We also manually rename some confusing results from this method.

```
original_names <- colnames(coffee_clean)
tidy_names <- gsub(" ", "_", original_names)
tidy_names <- tolower(tidy_names)
tidy_names <- gsub("[[:punct:]]&&[^_]", "", tidy_names)

colnames(coffee_clean) <- tidy_names

coffee_clean <- coffee_clean |>
  rename(
    age = "what_is_your_age?",
    cups_of_coffee_per_day = "how_many_cups_of_coffee_do_you_typically_drink_per_day?",
    how_else_at_home = "how_else_do_you_brew_coffee_at_home?",
    where_else_purchase_coffee = "where_else_do_you_purchase_coffee?",
    favorite_coffee_drink = "what_is_your_favorite_coffee_drink?",
    favorite_coffee = "please_specify_what_your_favorite_coffee_drink_is",
    prefer_between_abc = "between_coffee_a,_coffee_b,_and_coffee_c_which_did_you_prefer?",
    other_flavoring = "what_other_flavoring_do_you_use?",
    best_described_before = "before_today's_tasting,_which_of_the_following_best_described_wh
    like_coffee = "how_strong_do_you_like_your_coffee?",
    roast_level = "what_roast_level_of_coffee_do_you_prefer?",
    caffeine = "how_much_caffeine_do_you_like_in_your_coffee?",
    own_coffee_expertise = "lastly,_how_would_you_rate_your_own_coffee_expertise?",
    prefer_between_ad = "between_coffee_a_and_coffee_d,_which_did_you_prefer?",
```

```
    favorite_overall_coffee = "lastly,_what_was_your_favorite_overall_coffee?",
    time_spent_on_equipment = "approximately_how_much_have_you_spent_on_coffee_equipment_in_t
    good_value_equipment = "do_you_feel_like_you're_getting_good_value_for_your_money_with_re
  )

colnames(coffee_clean) <- sapply(colnames(coffee_clean), function(name) {
  if (grepl("where_do_you_typically_drink_coffee", name)) {
    name <- gsub("where_do_you_typically_drink_coffee\\?_\\((.*)\\)", "drink_\\1", name)
  } else if (grepl("how_do_you_brew_coffee_at_home", name)) {
    name <- gsub("how_do_you_brew_coffee_at_home\\?_\\((.*)\\)", "at_home_\\1", name)
  } else if (grepl("on_the_go,_where_do_you_typically_purchase_coffee", name)) {
    name <- gsub("on_the_go,_where_do_you_typically_purchase_coffee\\?_\\((.*)\\)", "purchase
  } else if (grepl("do_you_usually_add_anything_to_your_coffee", name)) {
    name <- gsub("do_you_usually_add_anything_to_your_coffee\\?_\\((.*)\\)", "add_to_\\1", na
  } else if (grepl("what_kind_of_dairy_do_you_add", name)) {
    name <- gsub("what_kind_of_dairy_do_you_add\\?_\\((.*)\\)", "dairy_add_\\1", name)
  } else if (grepl("what_kind_of_sugar_or_sweetener_do_you_add", name)) {
    name <- gsub("what_kind_of_sugar_or_sweetener_do_you_add\\?_\\((.*)\\)", "sugar_sweetener
  } else if (grepl("why_do_you_drink_coffee", name)) {
    name <- gsub("why_do_you_drink_coffee\\?_\\((.*)\\)", "reason_\\1", name)
  }
  name
}
)

#manually changing some more confusing names
coffee_clean_2 <- coffee_clean |>
  rename(at_home_coffee_brewing_machine = `at_home_coffee_brewing_machine_(e.g._mr._coffee)`
         at_home_pod_or_capsule_machine = `at_home_pod/capsule_machine_(e.g._keurig/nespresso
         at_home_coffee_extract = `at_home_coffee_extract_(e.g._cometeer)`,
         purchase_national_chain = `purchase_national_chain_(e.g._starbucks,_dunkin)`,
         add_to_none = `add_to_no_-_just_black`,
         add_to_milk = `add_to_milk,_dairy_alternative,_or_coffee_creamer`,
         sugar_sweetener_add_artificial_sweeteners = `sugar_sweetener_add_artificial_sweetene
         sugar_sweetener_add_raw_sugar= `sugar_sweetener_add_raw_sugar_(turbinado)`,
         where_work = `do_you_work_from_home_or_in_person?`,
         monthly_coffee_cost = `in_total,_much_money_do_you_typically_spend_on_coffee_in_a_mo
         like_taste = `do_you_like_the_taste_of_coffee?`,
         know_where_coffee_from = `do_you_know_where_your_coffee_comes_from?`,
         most_pay = `what_is_the_most_you've_ever_paid_for_a_cup_of_coffee?`,
         most_willing_pay = `what_is_the_most_you'd_ever_be_willing_to_pay_for_a_cup_of_coffe
         good_value_money = `do_you_feel_like_you're_getting_good_value_for_your_money_when_y
```

```
  mutate(`what_kind_of_dairy_do_you_add?` = NULL)

print(colnames(coffee_clean_2))
```

```
 [1] "submission_id"
 [2] "age"
 [3] "cups_of_coffee_per_day"
 [4] "drink_at_home"
 [5] "drink_at_the_office"
 [6] "drink_on_the_go"
 [7] "drink_at_a_cafe"
 [8] "drink_none_of_these"
 [9] "at_home_pour_over"
[10] "at_home_french_press"
[11] "at_home_espresso"
[12] "at_home_coffee_brewing_machine"
[13] "at_home_pod_or_capsule_machine"
[14] "at_home_instant_coffee"
[15] "at_home_bean-to-cup_machine"
[16] "at_home_cold_brew"
[17] "at_home_coffee_extract"
[18] "at_home_other"
[19] "how_else_at_home"
[20] "purchase_national_chain"
[21] "purchase_local_cafe"
[22] "purchase_drive-thru"
[23] "purchase_specialty_coffee_shop"
[24] "purchase_deli_or_supermarket"
[25] "purchase_other"
[26] "where_else_purchase_coffee"
[27] "favorite_coffee_drink"
[28] "favorite_coffee"
[29] "add_to_none"
[30] "add_to_milk"
[31] "add_to_sugar_or_sweetener"
[32] "add_to_flavor_syrup"
[33] "add_to_other"
[34] "what_else_do_you_add_to_your_coffee?"
[35] "dairy_add_whole_milk"
[36] "dairy_add_skim_milk"
[37] "dairy_add_half_and_half"
[38] "dairy_add_coffee_creamer"
```

```
[39] "dairy_add_flavored_coffee_creamer"
[40] "dairy_add_oat_milk"
[41] "dairy_add_almond_milk"
[42] "dairy_add_soy_milk"
[43] "dairy_add_other"
[44] "sugar_sweetener_add_granulated_sugar"
[45] "sugar_sweetener_add_artificial_sweeteners"
[46] "sugar_sweetener_add_honey"
[47] "sugar_sweetener_add_maple_syrup"
[48] "sugar_sweetener_add_stevia"
[49] "sugar_sweetener_add_agave_nectar"
[50] "sugar_sweetener_add_brown_sugar"
[51] "sugar_sweetener_add_raw_sugar"
[52] "other_flavoring"
[53] "best_described_before"
[54] "like_coffee"
[55] "roast_level"
[56] "caffeine"
[57] "own_coffee_expertise"
[58] "coffee_a_-_bitterness"
[59] "coffee_a_-_acidity"
[60] "coffee_a_-_personal_preference"
[61] "coffee_a_-_notes"
[62] "coffee_b_-_bitterness"
[63] "coffee_b_-_acidity"
[64] "coffee_b_-_personal_preference"
[65] "coffee_b_-_notes"
[66] "coffee_c_-_bitterness"
[67] "coffee_c_-_acidity"
[68] "coffee_c_-_personal_preference"
[69] "coffee_c_-_notes"
[70] "coffee_d_-_bitterness"
[71] "coffee_d_-_acidity"
[72] "coffee_d_-_personal_preference"
[73] "coffee_d_-_notes"
[74] "prefer_between_abc"
[75] "prefer_between_ad"
[76] "favorite_overall_coffee"
[77] "where_work"
[78] "monthly_coffee_cost"
[79] "reason_it_tastes_good"
[80] "reason_i_need_the_caffeine"
[81] "reason_i_need_the_ritual"
```

```
[82] "reason_it_makes_me_go_to_the_bathroom"
[83] "reason_other"
[84] "other_reason_for_drinking_coffee"
[85] "like_taste"
[86] "know_where_coffee_from"
[87] "most_pay"
[88] "most_willing_pay"
[89] "good_value_money"
[90] "time_spent_on_equipment"
[91] "good_value_equipment"
[92] "gender"
[93] "education_level"
[94] "ethnicity/race"
[95] "ethnicity/race_(please_specify)"
[96] "employment_status"
[97] "number_of_children"
[98] "political_affiliation"
```

4. After renaming our columns, we noticed some of them work nicely as categorical factors. This section goes through and modify them to be factors in a logical order.

```r
#add category
coffee_clean_factors <- coffee_clean_2 |>
  mutate(age = factor(age)) |>
  mutate(age = fct_relevel(age, c("<18 years old",
                                  "18-24 years old",
                                  "25-34 years old",
                                  "35-44 years old",
                                  "45-54 years old",
                                  "55-64 years old",
                                  ">65 years old")))|>

  mutate(monthly_coffee_cost = factor(monthly_coffee_cost),
         monthly_coffee_cost = fct_relevel(monthly_coffee_cost, c(
    "<$20",
    "$20-$40",
    "$40-$60",
    "$60-$80",
    "$80-$100",
    ">$100"))) |>
  mutate(across(like_taste:political_affiliation, factor)) |>
  mutate(across(like_coffee:caffeine, factor))
  # mutate(`what_is_the_most_you've_ever_paid_for_a_cup_of_coffee?` = fct_relevel(
```

```
#   "Less than $2",
#   "$2-$4",
#   "$4-$6",
#   "$6-$8",
#   "$8-$10",
#   "$10-$15",
#   "$15-$20",
#   "More than $20"
# )) |>
#   mutate(`what_is_the_most_you'd_ever_be_willing_to_pay_for_a_cup_of_coffee?`) = fct_rele
#   "Less than $2",
#   "$2-$4",
#   "$4-$6",
#   "$6-$8",
#   "$8-$10",
#   "$10-$15",
#   "$15-$20",
#   "More than $20"
# )
```

Data description

Have an initial draft of your data description section. Your data description should be about your analysis-ready data.

Data limitations

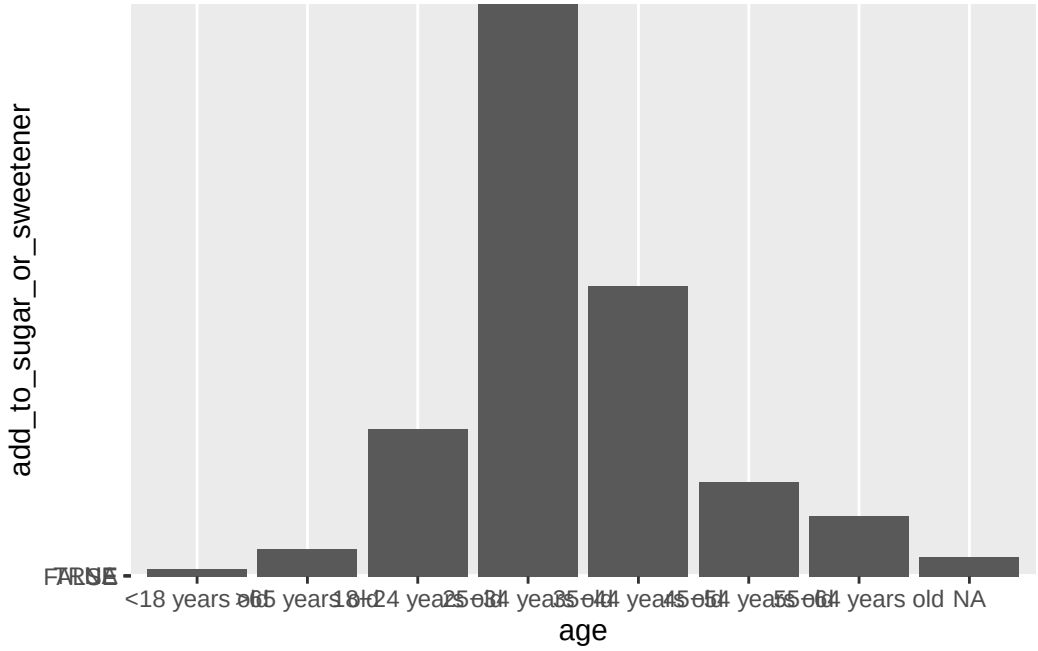
Identify any potential problems with your dataset.

There are many NA values across the dataset.

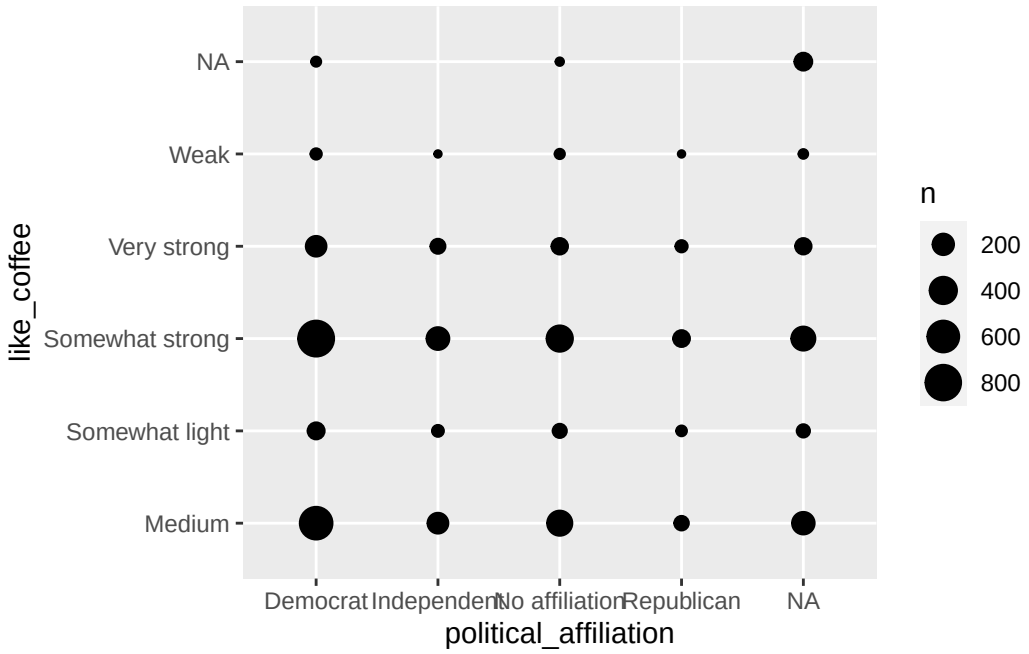
Exploratory data analysis

Perform an (initial) exploratory data analysis.

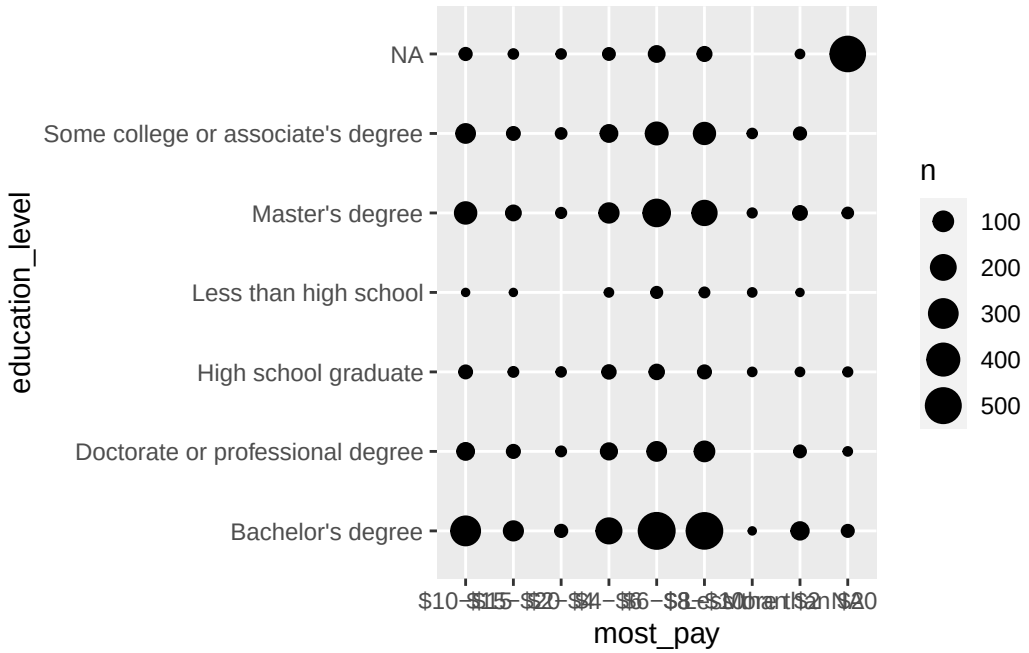
```r
#Joice graphs
#sweetener by age
coffee_clean_2 |>
  ggplot(aes(x = age, y = add_to_sugar_or_sweetener)) +
  geom_col()
```



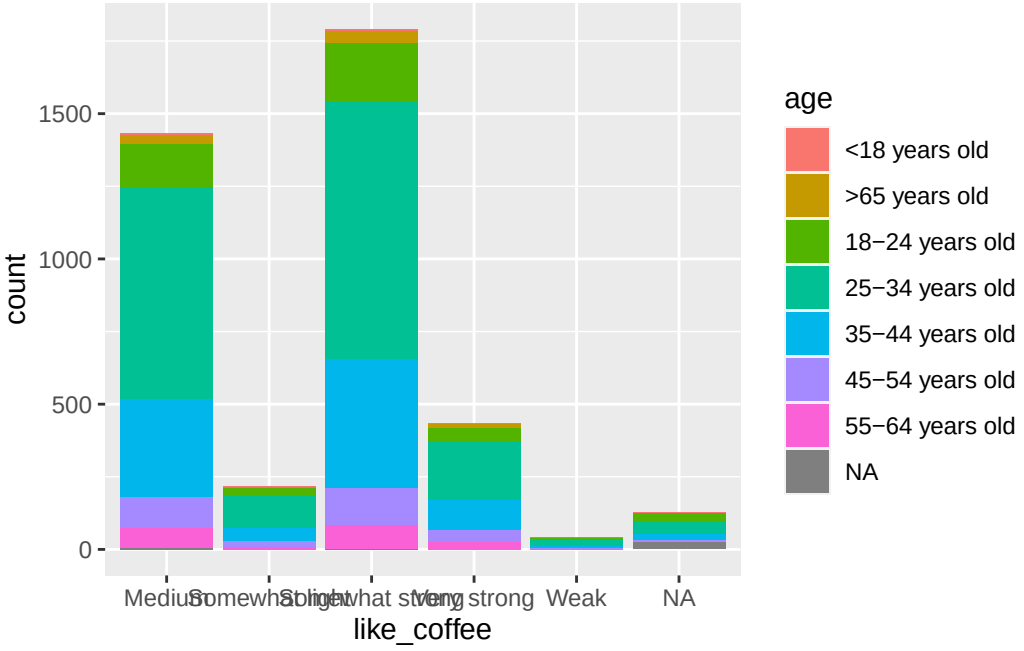

```r
#coffee payment political affiliation
coffee_clean_2 |>
  ggplot(aes(x = political_affiliation, y = like_coffee)) +
  geom_count()
```



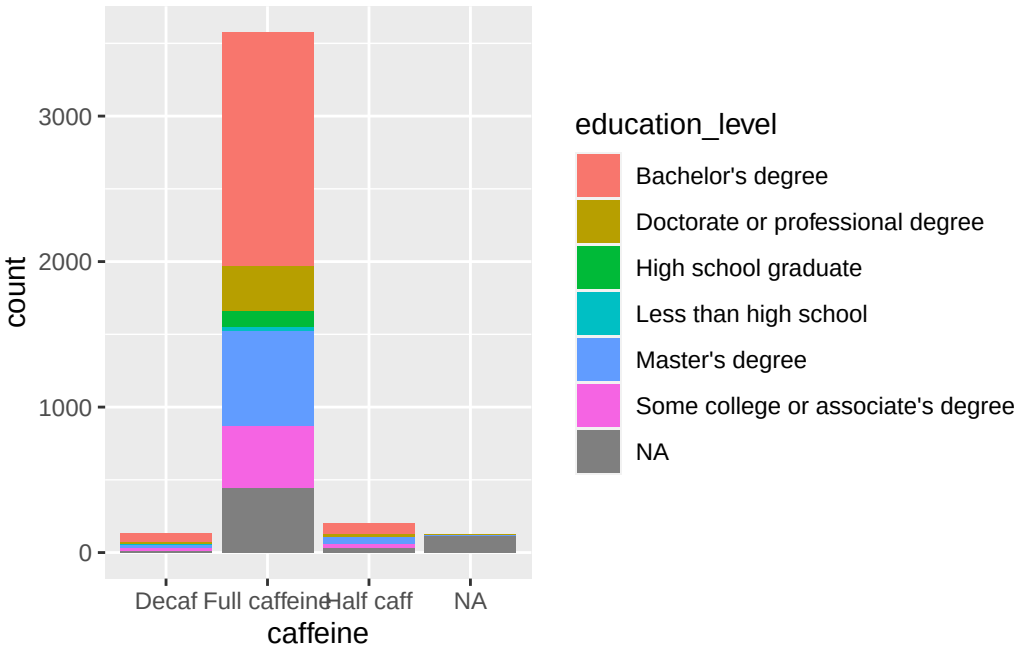

```
#coffee expected and actual pay
coffee_clean_2 |>
  ggplot(aes(x = most_pay, y = education_level)) +
  geom_count()
```



```
#coffee like or not by age
coffee_clean_2 |>
  ggplot(aes(x = like_coffee, fill = age)) +
  geom_bar()
```




```
#coffee like or not by age
coffee_clean_2 |>
  ggplot(aes(x = caffeine, fill = education_level)) +
  geom_bar()
```



Questions for reviewers

List specific questions for your peer reviewers and project mentor to answer in giving you feedback on this phase.