Project 3

This is the dataset you will be working with:

```
food <- readr::read_csv("https://wilkelab.org/DSC385/datasets/food_coded.csv")
food
```

```
## # A tibble: 125 x 61
##    GPA   Gender breakfast calor~1 calor~2 calor~3 coffee comfo~4 comfo~5 comfo~6
##    <chr>  <dbl>     <dbl>   <dbl>   <dbl>   <dbl>  <dbl> <chr>   <chr>     <dbl>
##  1 2.4        2         1     430     NaN     315      1 none    we don~       9
##  2 3.654      1         1     610       3     420      2 chocol~ Stress~       1
##  3 3.3        1         1     720       4     420      2 frozen~ stress~       1
##  4 3.2        1         1     430       3     420      2 Pizza,~ Boredom       2
##  5 3.5        1         1     720       2     420      2 Ice cr~ Stress~       1
##  6 2.25       1         1     610       3     980      2 Candy,~ None, ~       4
##  7 3.8        2         1     610       3     420      2 Chocol~ stress~       1
##  8 3.3        1         1     720       3     420      1 Ice cr~ I eat ~       1
##  9 3.3        1         1     430     NaN     420      1 Donuts~ Boredom       2
## 10 3.3        1         1     430       3     315      2 Mac an~ Stress~       1
## # ... with 115 more rows, 51 more variables: cook <dbl>,
## #   comfort_food_reasons_coded...12 <dbl>, cuisine <dbl>, diet_current <chr>,
## #   diet_current_coded <dbl>, drink <dbl>, eating_changes <chr>,
## #   eating_changes_coded <dbl>, eating_changes_coded1 <dbl>, eating_out <dbl>,
## #   employment <dbl>, ethnic_food <dbl>, exercise <dbl>,
## #   father_education <dbl>, father_profession <chr>, fav_cuisine <chr>,
## #   fav_cuisine_coded <dbl>, fav_food <dbl>, food_childhood <chr>, ...
```

Question: Is GPA related to student income, the father's educational level, or the student's perception of what an ideal diet is?

Introduction:

To answer the question: whether a student's GPA is related to Income, Father's education level, student's perception of ideal diet, the actual `food` dataset is used. The dataset contains 125 records, 61 columns among which 4 columns are extracted and stored in a filtered dataset named `food_filter` which contains the following columns: The Grade Point Average of the student (column `GPA`), student income (column `income`), father's education level (column `father_education`) and student's perception of ideal diet (column `ideal_diet_coded`).

From an initial scan of the data using `summary()` and `table()`, it was observed that GPA column is string type and has 5 invalid values, `father_education` column uses numbers 1-5 to annotate level of education and has 1 invalid value, `income` column uses numbers 1-6 to depict income ranges and has 1 NA value and `ideal_diet_coded` column uses numbers 1-8 to depict different categories of students' diet perception. The last three columns are very hard to be interpreted by any random person because they are coded as integers which has to be referenced manually in the detailed data dictionary. *Hence, columns have to be recoded to have a clean dataset before plotting the relationship between GPA and other 3 metrics.*

Approach:

After filtering the the 4 columns from the `food` dataset and getting the `food_filter` subset, the GPA column is converted into numeric value using the `as.numeric()` function and all invalid values are thus replaced by **NA**. `income`, `father_education`, `ideal_diet_coded` columns are recoded using `case_When()` function while referencing the appropriate code with the data dictionary. While using case_when, a number is added in front of each category so that it will be ordered automatically in plotting. **NA values have been dropped from the final dataset, thereby reducing the count from 125 records to 119 records**. After recoding the data, the columns are checked by using `summary()` and `table()`.

Our approach is to show the distributions of GPA plotted against `income`, `father_education` and `ideal_diet_coded` separately. This can be best visualized using **geom_boxplot()** because it shows proper distribution for a category variable versus a continuous variable. Since the dataset is very small in number, other plots like violin, ridges might not be visually appropriate.

Analysis:

First we filter the dataset, then code the categorical variables as per data dictionary.

```
food_filter <-
  food |>
  # select 4 required columns for analysis
  select(GPA, father_education, income, ideal_diet_coded) |>
  mutate(GPA = as.numeric(GPA)) |> # converting GPA to numeric
  drop_na() # dropping 6 NA records

food_filter <-
  food_filter |>
  mutate(
    ideal_diet_coded = case_when(
      # converting ideal_diet_coded column into categorical
      ideal_diet_coded == 1 ~ "Portion control",
      ideal_diet_coded == 2 ~ "Adding veggies/fruit/eating healthier",
      ideal_diet_coded == 3 ~ "Balance",
      ideal_diet_coded == 4 ~ "Less sugar",
      ideal_diet_coded == 5 ~ "Home cooked/organic",
      ideal_diet_coded == 6 ~ "Current diet",
      ideal_diet_coded == 7 ~ "More protein",
      ideal_diet_coded == 8 ~ "Unclear"
    )
  ) |>
  mutate(
    father_education = case_when(
      # converting father_education column to describe father_education
      father_education == 1 ~ "1. Less than high school",
      father_education == 2 ~ "2. High school degree",
      father_education == 3 ~ "3. Some college degree",
      father_education == 4 ~ "4. College degree",
      father_education == 5 ~ "5. Graduate degree"
    )
  ) |>
  mutate(
    income = case_when(
      # converting income column to describe income as a categorical variable
      income == 1 ~ "1. Less than $15000",
```

```
      income == 2 ~ "2. $15001 to $30000",
      income == 3 ~ "3. $30001 to $50000",
      income == 4 ~ "4. $50000 to $70000",
      income == 5 ~ "5. $70001 to $100000",
      income == 6 ~ "6. Higher than $100000"
    )
  )
```

Now we check the summary and table of each of four columns as asked in the Question.

```
food_filter |>
  select(GPA) |> # printing summary of GPA after changing it to numeric
  summary()
```

```
##       GPA
##  Min.   :2.200
##  1st Qu.:3.200
##  Median :3.500
##  Mean   :3.418
##  3rd Qu.:3.700
##  Max.   :4.000
```

```
# Checking and printing the summary of each categorical variable after recoding

food_filter |>
  select(ideal_diet_coded) |>

  table(useNA = "ifany")
```

```
## ideal_diet_coded
## Adding veggies/fruit/eating healthier                              Balance
##                                    41                                   15
##                            Current diet                  Home cooked/organic
##                                    13                                   15
##                              Less sugar                         More protein
##                                     6                                   15
##                         Portion control                              Unclear
##                                    11                                    3
```

```
food_filter |>
  select(father_education) |>
  table(useNA = "ifany")
```

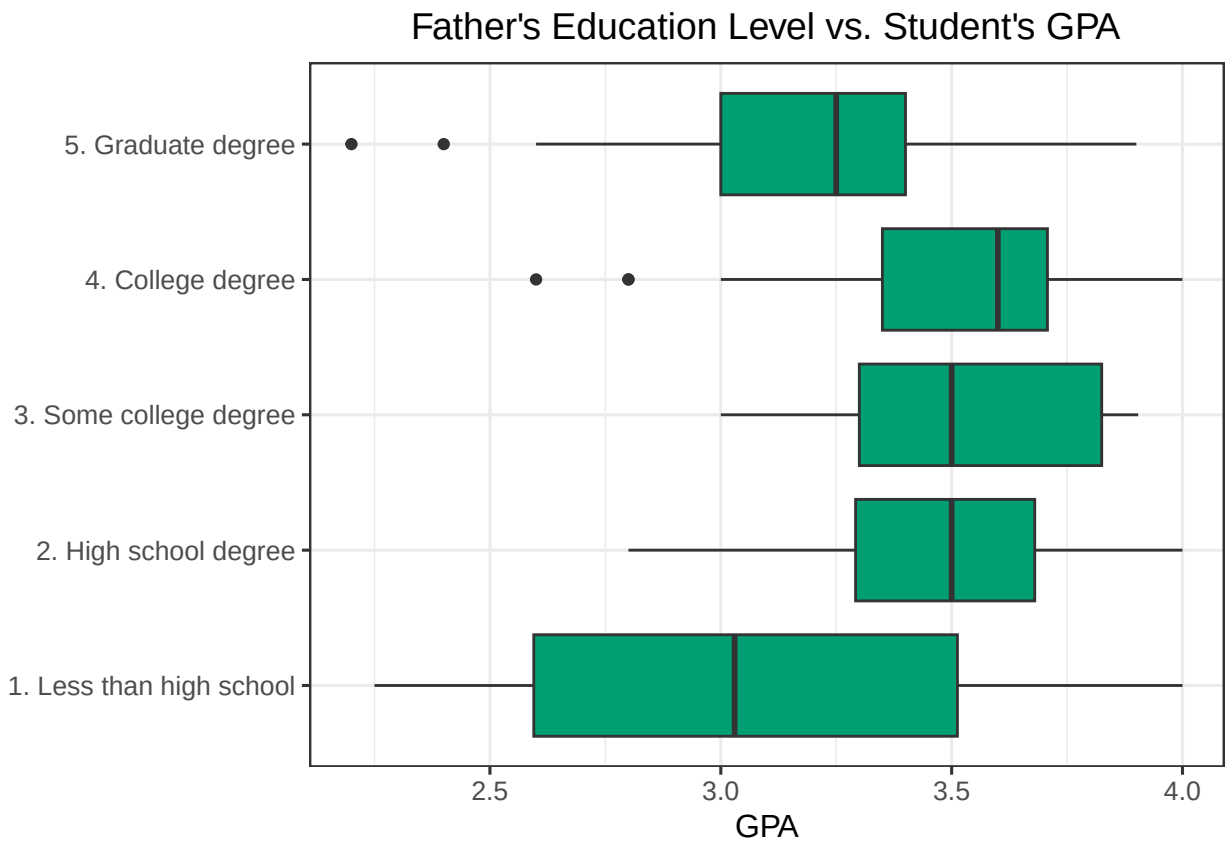
```
## father_education
## 1. Less than high school    2. High school degree   3. Some college degree
##                        4                       33                       12
##        4. College degree       5. Graduate degree
##                       44                       26
```

```
food_filter |>
  select(income) |>
  table(useNA = "ifany")
```

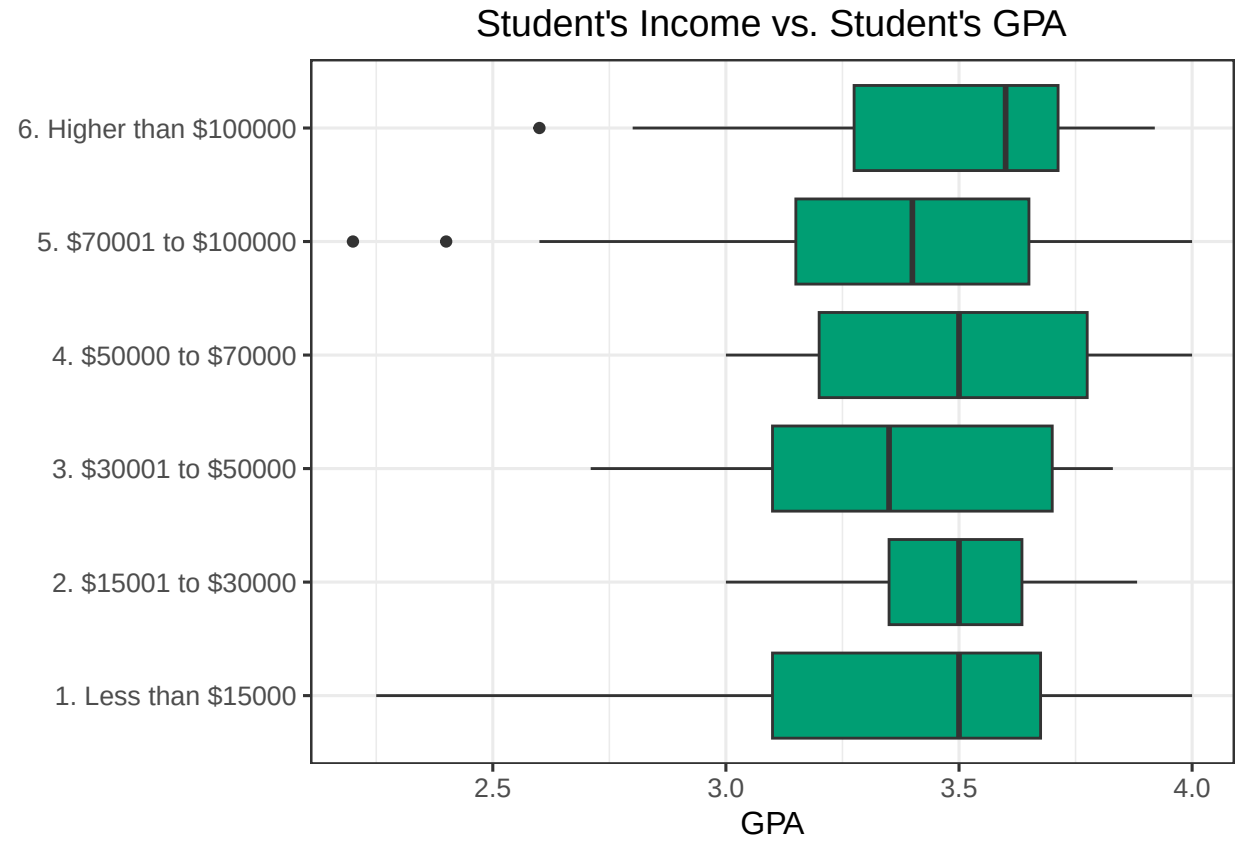
```
## income
##    1. Less than $15000    2. $15001 to $30000    3. $30001 to $50000
##                      6                      7                     17
##    4. $50000 to $70000   5. $70001 to $100000 6. Higher than $100000
##                     18                     31                     40
```

Now we plot the relationship between GPA and three other metrics in separate plots and discuss the results in the end.

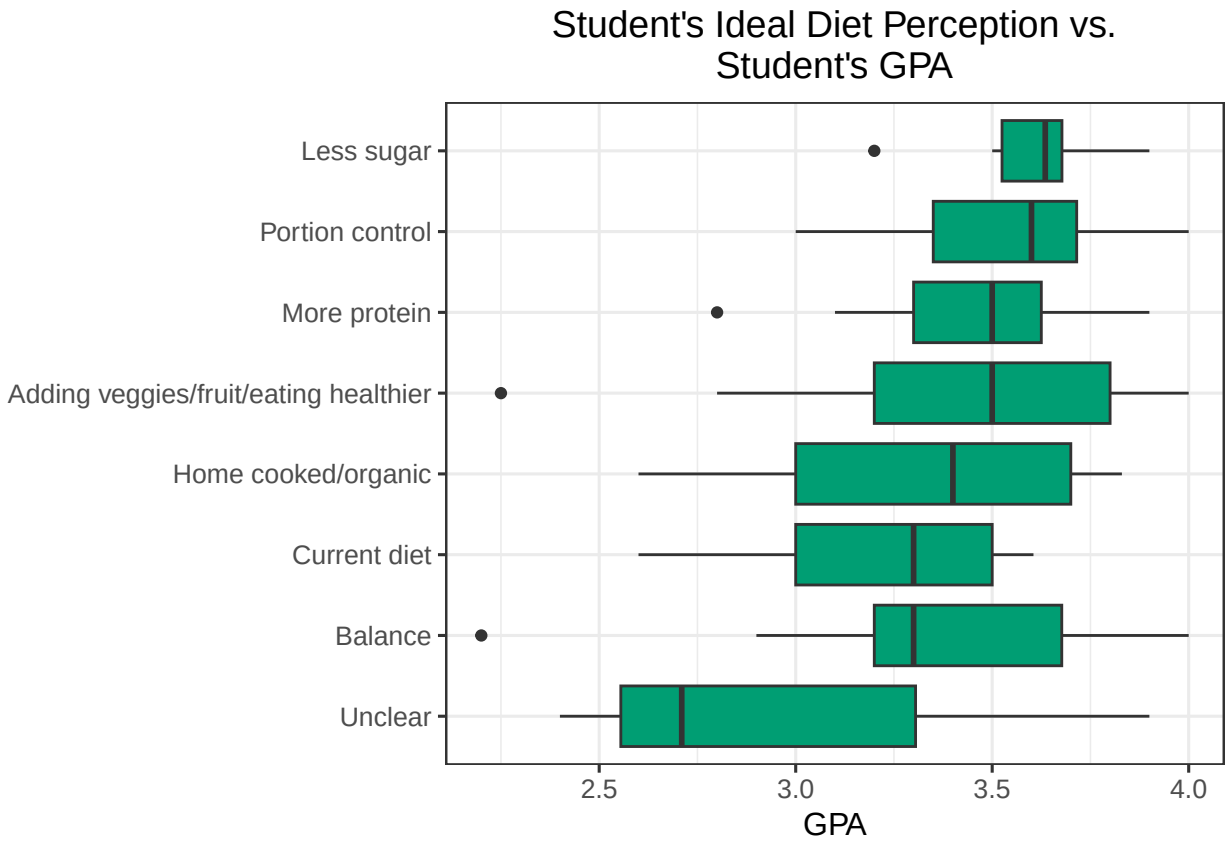
```
food_filter |>
  filter(!is.na(father_education)) |>
  # father_education column with NA values are not plotted
  ggplot(aes(GPA, father_education)) +
  geom_boxplot(fill = "#009E73") +
  scale_x_continuous(name = "GPA") +
  scale_y_discrete(name = NULL) +
  ggtitle("Father's Education Level vs. Student's GPA") +
  theme_bw(12) +
  theme(plot.title = element_text(hjust = 0.5))
```



```r
food_filter |>
  filter(!is.na(income)) |>
  # income column with NA values are not plotted
  ggplot(aes(GPA, income)) +
  geom_boxplot(fill = "#009E73") +
  scale_x_continuous(name = "GPA") +
  scale_y_discrete(name = NULL) +
  ggtitle("Student's Income vs. Student's GPA") +
  theme_bw(12) +
  theme(plot.title = element_text(hjust = 0.5))
```




```r
food_filter |>
  mutate(ideal_diet_coded = fct_reorder(ideal_diet_coded, GPA, na.rm = TRUE)) |>
  ggplot(aes(GPA, ideal_diet_coded)) +
  geom_boxplot(fill = "#009E73") +
  scale_x_continuous(name = "GPA") +
  scale_y_discrete(name = NULL) +
  ggtitle("Student's Ideal Diet Perception vs.\nStudent's GPA") +
  theme_bw(12) +
  theme(plot.title = element_text(hjust = 0.5))
```



Discussion:

In the *first plot* "Father's Education Level vs. Student's GPA", it is clearly evident that if the father has done some education above high school the majority of the students' GPA fall above 3. The "less than highschool" category has the widest GPA distribution. In the *second plot* "Student's Income vs. Student's GPA", there seems to be no relationship between these two variables as the distribution is uniform across all income ranges.

In the *third plot*, "Student's Ideal Diet Perception vs. Student's GPA" the median GPA across each boxplot seems to be increasing in the order of *unclear, balance, current diet, home cooked, adding veggies, more protein, portion control, less sugar* categories. Thus we can observe that if a student has a proper diet, they tend to have a better GPA.

Thus, the final answer to question is: There seems to be no relationship between GPA and student Income. However, there is a mild observation that if a father has education less than highschool, then their GPA tends to majorly fall in the range 2.6 to 3.5. But most importantly, the student's perception of ideal diet seems to have a trend. The better diet perception a student has, the higher his/her GPA.