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 calories_chicken calories_day calories_scone coffee
##    <chr>  <dbl>     <dbl>            <dbl>        <dbl>          <dbl>  <dbl>
##  1 2.4        2         1              430          NaN            315      1
##  2 3.654      1         1              610            3            420      2
##  3 3.3        1         1              720            4            420      2
##  4 3.2        1         1              430            3            420      2
##  5 3.5        1         1              720            2            420      2
##  6 2.25       1         1              610            3            980      2
##  7 3.8        2         1              610            3            420      2
##  8 3.3        1         1              720            3            420      1
##  9 3.3        1         1              430          NaN            420      1
## 10 3.3        1         1              430            3            315      2
## # ... with 115 more rows, and 54 more variables: comfort_food <chr>,
## #   comfort_food_reasons <chr>, comfort_food_reasons_coded...10 <dbl>,
## #   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>, ...
```

A detailed data dictionary for this dataset is available here. (https://wilkelab.org/DSC385/datasets/food_codebook.pdf) The dataset was originally downloaded from Kaggle, and you can find additional information about the dataset here. (https://www.kaggle.com/borapajo/food-choices/version/5)

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

To answer this question, first prepare a cleaned dataset that contains only the four relevant data columns, properly cleaned so that numerical values are stored as numbers and categorical values are represented by humanly readable words or phrases. For categorical variables with an inherent order, make sure the levels are in the correct order.

In your introduction, carefully describe each of the four relevant data columns. In your analysis, provide a summary of each of the four columns, using `summary()` for numerical variables and `table()` for categorical variables.

Then, make one visualization each for student income, father's educational level, and ideal diet, and answer the question separately for each visualization. The three visualizations can be of the same type.

Hints:

1. Use `case_when()` to recode categorical variables.

2. Use `fct_relevel()` to arrange categorical variables in the right order.
3. Use `as.numeric()` to convert character strings into numerical values. It is fine to ignore warnings about `NA` s introduced by coercion.
4. `NaN` stands for Not a Number and can be treated like `NA` . You do not need to replace `NaN` with `NA` .
5. When using `table()` , provide the argument `useNA = "ifany"` to make sure missing values are counted: `table(..., useNA = "ifany")` .

Introduction: The dataset used in the study is the `food_coded` dataset provided by *Kaggle*. The data consists of survey answers from the students of *Mercyhurst University* on various preferences of food as well as some other general information on the student such as `GPA` and `weight` . For this analysis, we will use the `income` , `father_education` , and `ideal_diet_coded` data columns to predict the students `GPA` . The `GPA` is the standard four point system used at most universities. The `income` data is coded from 1 to 6 representing ranges of income. The `father_education` data is coded from 1 to 5 indicating what level of education the student's father has earned such as high school diploma or college degree. The `ideal_diet_coded` data is coded from 1 to 8 with various answers from students on what they consider to be important to an ideal diet in the context of their current diet. Answers to this survey range from portion control to adding veggies. The data is cleaned, isolated, and displayed below.

```
food <- food %>%
  mutate(GPA_numeric = as.numeric(GPA),
         income_readable = case_when(
           income == 1 ~ "less than $15,000",
           income == 2 ~ "$15,001 to $30,000",
           income == 3 ~ "$30,001 to $50,000",
           income == 4 ~ "$50,001 to $70,000",
           income == 5 ~ "$70,001 to $100,000",
           income == 6 ~ "higher than $100,000",
           TRUE ~ NA_character_
           ),
         father_education_readable = case_when(
           father_education == 1 ~ "less than high school",
           father_education == 2 ~ "high school degree",
           father_education == 3 ~ "some college degree",
           father_education == 4 ~ "college degree",
           father_education == 5 ~ "graduate degree",
           TRUE ~ NA_character_
           ),
         ideal_diet_readable = case_when(
           ideal_diet_coded == 1 ~ "portion control",
           ideal_diet_coded == 2 ~ "eating healthier food",
           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",
           TRUE ~ NA_character_
           )
         )
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
summary(food$GPA_numeric)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   2.200   3.200   3.500   3.416   3.700   4.000       5
```

```
as.data.frame(table(food$income_readable, useNA = "ifany"))
```

```
##                   Var1 Freq
## 1   $15,001 to $30,000    7
## 2   $30,001 to $50,000   17
## 3   $50,001 to $70,000   20
## 4  $70,001 to $100,000   33
## 5 higher than $100,000   41
## 6    less than $15,000    6
## 7                 <NA>    1
```

```
as.data.frame(table(food$father_education_readable, useNA = "ifany"))
```

```
##                    Var1 Freq
## 1        college degree   46
## 2       graduate degree   28
## 3    high school degree   34
## 4 less than high school    4
## 5   some college degree   12
## 6                  <NA>    1
```

```
as.data.frame(table(food$ideal_diet_readable, useNA = "ifany"))
```

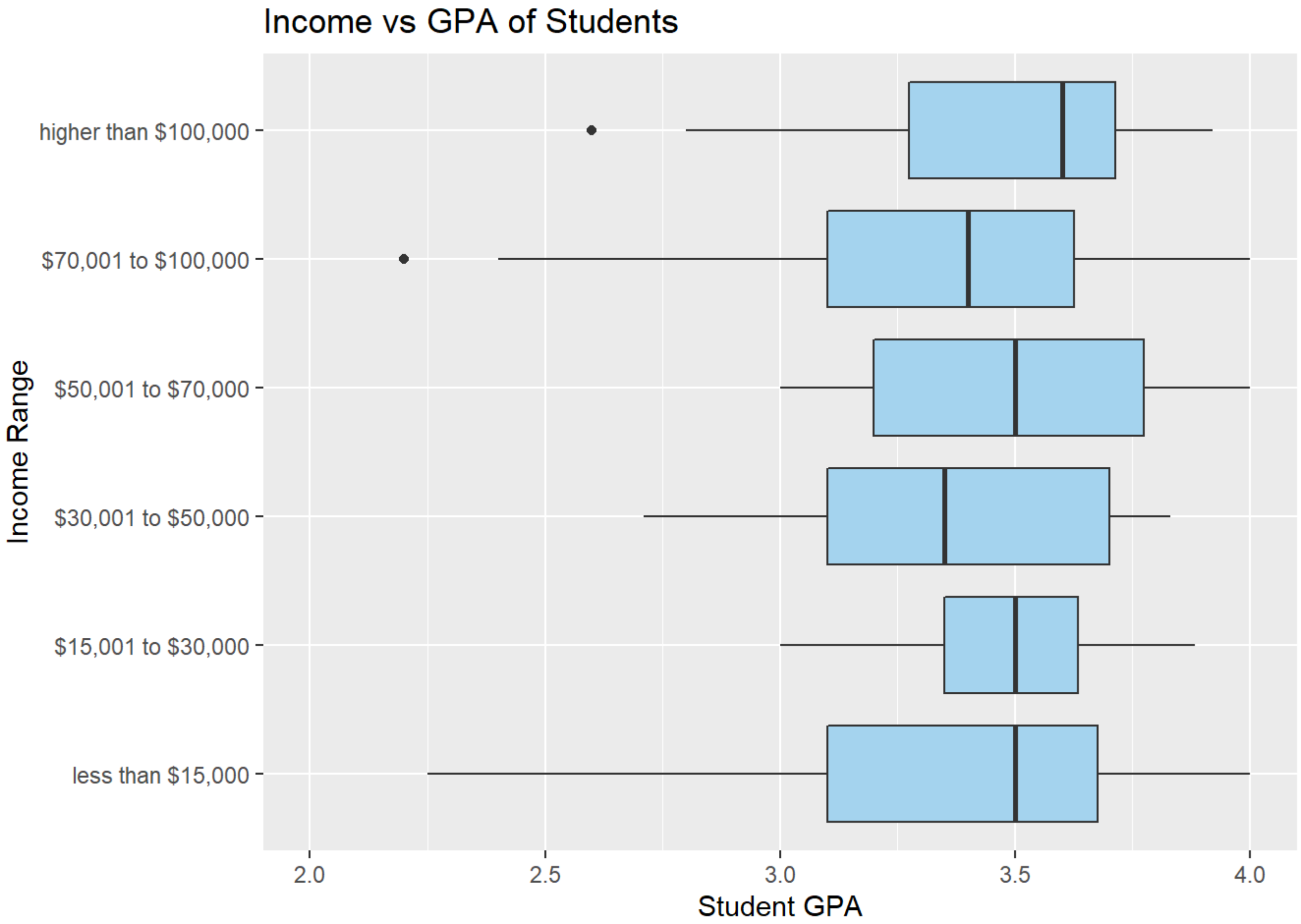
```
##                    Var1 Freq
## 1               balance   17
## 2          current diet   13
## 3 eating healthier food   44
## 4   home cooked/organic   15
## 5            less sugar    6
## 6          more protein   16
## 7       portion control   11
## 8               unclear    3
```

Approach: Because the independent data given is in all cases categorical, we will use a series of boxplots. The boxplots will be separated by the independent variable and aligned on the same axis to make clear comparisons of the student's GPA. Boxplots are also beneficial in that they show the medians of the student's GPAs for each category; this allows us to track a trend if one can be seen. The GPA axis of the plots will limit from 2 to 4 to keep them alike and not skew the scale across each. It will also give context with relation to the bottom and top performers.

Analysis:

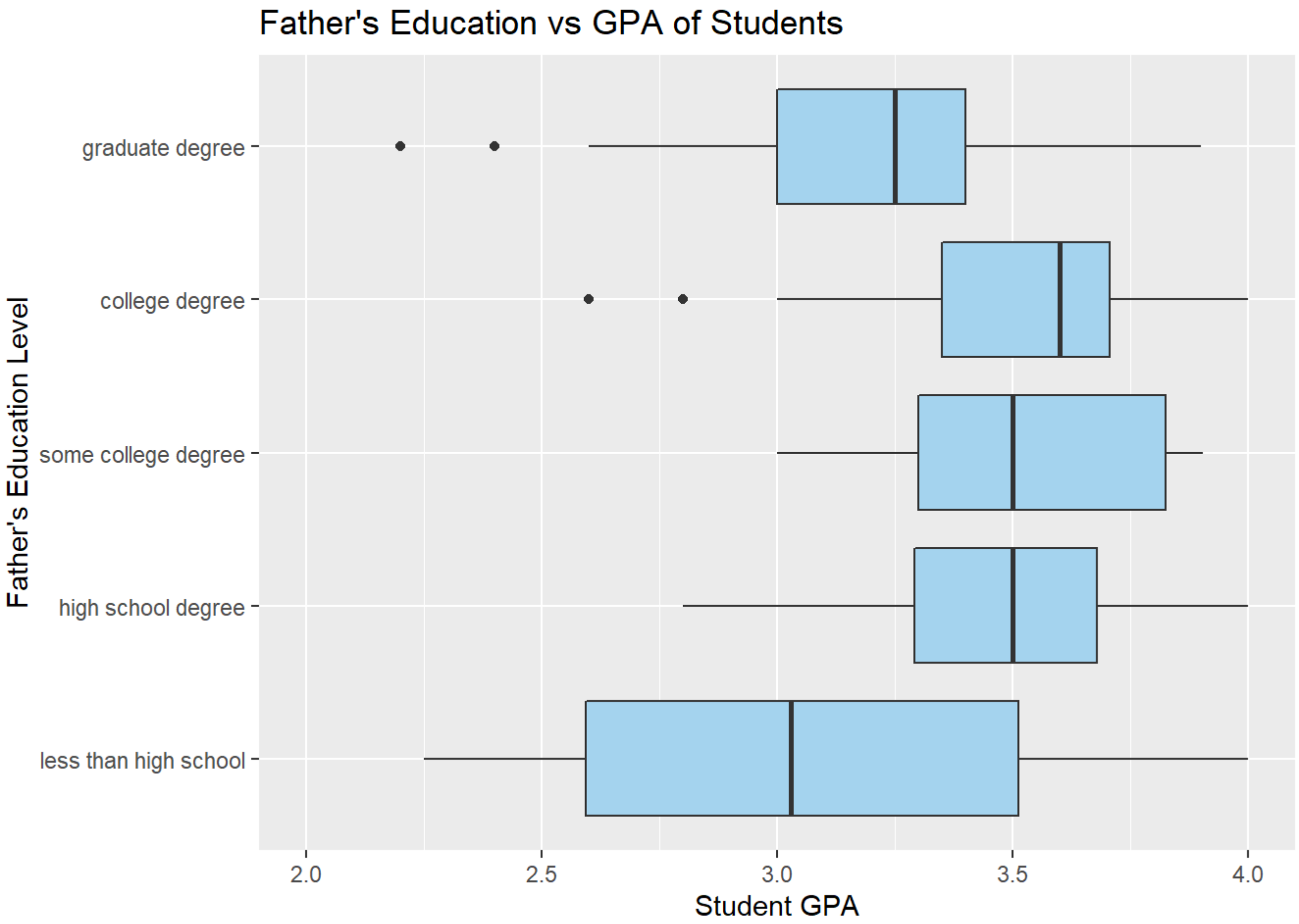
Here we filter out all GPAs that are invalid (`N/A`) and we filter out all income values that are invalid (`N/A`). We plot the boxplots in order of the income labels for ease of readability.

```
food %>%
  filter(
    !is.na(GPA_numeric),
    !is.na(income_readable)
    ) %>%
  ggplot(aes(GPA_numeric, fct_reorder(income_readable, income))) +
  geom_boxplot(fill = "lightskyblue2") +
  scale_x_continuous(
    name = "Student GPA",
    limits = c(2, 4),
    breaks = c(2, 2.5, 3, 3.5, 4)
  ) +
  labs(
    title = "Income vs GPA of Students",
    y = "Income Range")
```



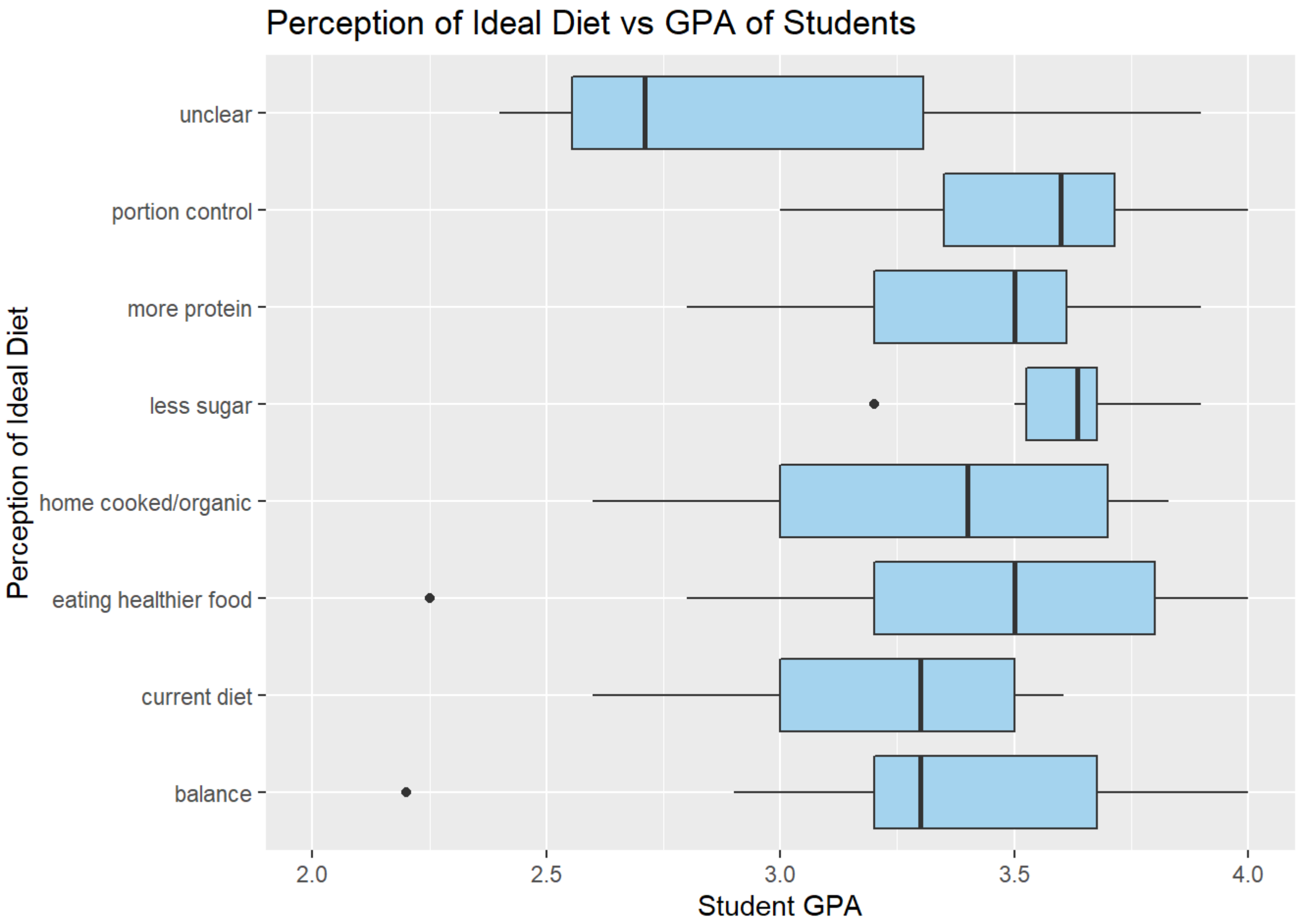
Here we filter out all GPAs that are invalid (`N/A`) and we filter out all father education values that are invalid (`N/A`). We plot the boxplots in order of the father’s income labels for ease of readability.

```
food %>%
  filter(
    !is.na(GPA_numeric),
    !is.na(father_education_readable)
    ) %>%
  ggplot(aes(GPA_numeric, fct_reorder(father_education_readable, father_education))) +
  geom_boxplot(fill = "lightskyblue2") +
  scale_x_continuous(
    name = "Student GPA",
    limits = c(2, 4),
    breaks = c(2, 2.5, 3, 3.5, 4)
  ) +
  labs(
    title = "Father's Education vs GPA of Students",
    y = "Father's Education Level")
```



Here we filter out all GPAs that are invalid (`N/A`) and we filter out all ideal diet values that are invalid (`N/A`). Because there is no inherent order to the independent variable, the boxplots are not ordered.

```
food %>%
  filter(
    !is.na(GPA_numeric),
    !is.na(ideal_diet_readable)
    ) %>%
  ggplot(aes(GPA_numeric, ideal_diet_readable)) +
  geom_boxplot(fill = "lightskyblue2") +
  scale_x_continuous(
    name = "Student GPA",
    limits = c(2, 4),
    breaks = c(2, 2.5, 3, 3.5, 4)
  ) +
  labs(
    title = "Perception of Ideal Diet vs GPA of Students",
    y = "Perception of Ideal Diet")
```



Discussion:

Income vs GPA of Students:

From the `Income vs GPA of Students` plot above, we see that there does not appear to be any significant correlation between the student's income and their GPA. The medians of each range are nearly the same across all categories. There are perhaps a few under-performers in the poorest and richest categories. We could presume this to be either students with difficult life struggles and spoiled students who don't have the need or understand the value of education respectively, but there is not much data to assume this.

Father’s Education vs GPA of Students:

The education of the father of the student does appear to suggest some correlation to the students GPA by the `Father's Education vs GPA of Students` plot above. The students who were fathered by a high school drop out are statistically more likely to perform worse than their peers. Father’s with a graduate degree are also more likely to have a son that does not perform as well in school. We could presume that a high school drop out will not be a responsible father or would not value education.

Perception of Ideal Diet vs GPA of Students:

As shown from the above plot called `Perception of Ideal Diet vs GPA of Students` , we can see that most categories do not correlate to the student’s GPA. The medians are slightly more varied from the `Father's Education vs GPA of Students` plot, albeit, with one exception. The `unclear` category interestingly shows a sharp decrease in student performance. We could presume that students with no strong input are not very organized or responsible or forward thinking. We again do not have the necessary information to conclude any more than this, however.