Project 3

This is the dataset you will be working with:

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")` .

Data Cleansing

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

# Print first few rows
head(food)

# Get column names
colnames(food)

# Create new dataset "food_relevant" that only selects relevant columns
food_relevant <- select(food, GPA, income, father_education, ideal_diet_coded)

# Show summary of those columns
summary(food_relevant)

# Show shape of dataset
dim(food_relevant)
# 125 rows by 4 cols


sum(is.na(food_relevant$GPA))
# 0 NAs
sum(is.na(food_relevant$income))
# 1 NA
sum(is.na(food_relevant$father_education))
# 1 NA
sum(is.na(food_relevant$ideal_diet_coded))
# 0 NA
#Income & father_education each had 1 na

na.omit(food_relevant)
# 123 rows by 4 (2 rows omitted)

food_cleaned <-
  #Omit 2 nulls
  na.omit(food_relevant) |>
    mutate(
    #Convert GPA to numeric
    gpa = as.numeric(GPA),

    #Case When Income
    income = case_when(
      income == 1 ~ "< $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 ~ "> $100,000"
    ),

    father_education = 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"
    ),

    ideal_diet = case_when(
      ideal_diet_coded == 1 ~ "Portion Control",
      ideal_diet_coded == 2 ~ "Adding Veggies/Eating Healthier Food/Adding Fruit",
      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",
    )
  ) |>
select(gpa, income, father_education, ideal_diet)
summary(food_cleaned)
```

Introduction:

The dataset that will be used to answer the question is about food choices and preferences from responses of 126 college students from Mercyhurst University who participated in a survey. In order to determine if GPA is related to student income, the father's educational level, or the student's perception of what an ideal diet is, I will use the following columns:

1. GPA: The grade point average of the student
2. income: The amount of money the student earns
3. father_education: The level of education obtained by the student's father
4. ideal_diet_coded: Coded ideal diet from open-ended answer of the student

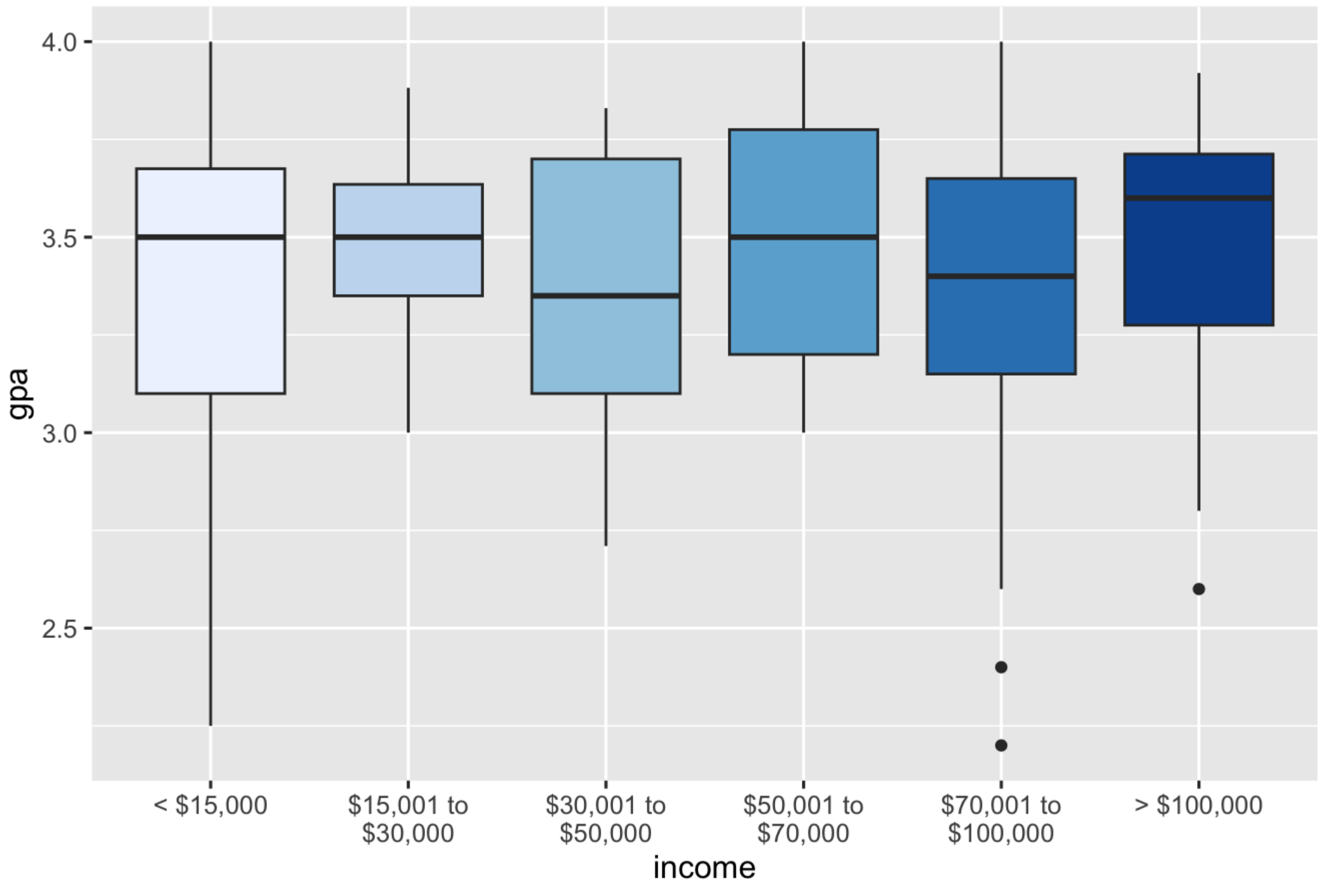
GPA will be my response variable, while the others will be the explanatory variables. As these are all categorical variables, I will also convert the numerical codes to human readable values that are easier to interpret. Lastly, I will remove all NAs as they are not relevant for this analysis.

Approach:

After only retaining the relevant columns and removing NAs from the analysis in the Introduction section mentioned above, since all explanatory variables (income, father_education & ideal_diet_coded) are categorical variables, I will utilize boxplots to analyze the relationship between each group within each field to determine if there is a relationship between the explanatory variable in question and GPA. For example, one might assume that the student's income may be positively correlated with their GPA.

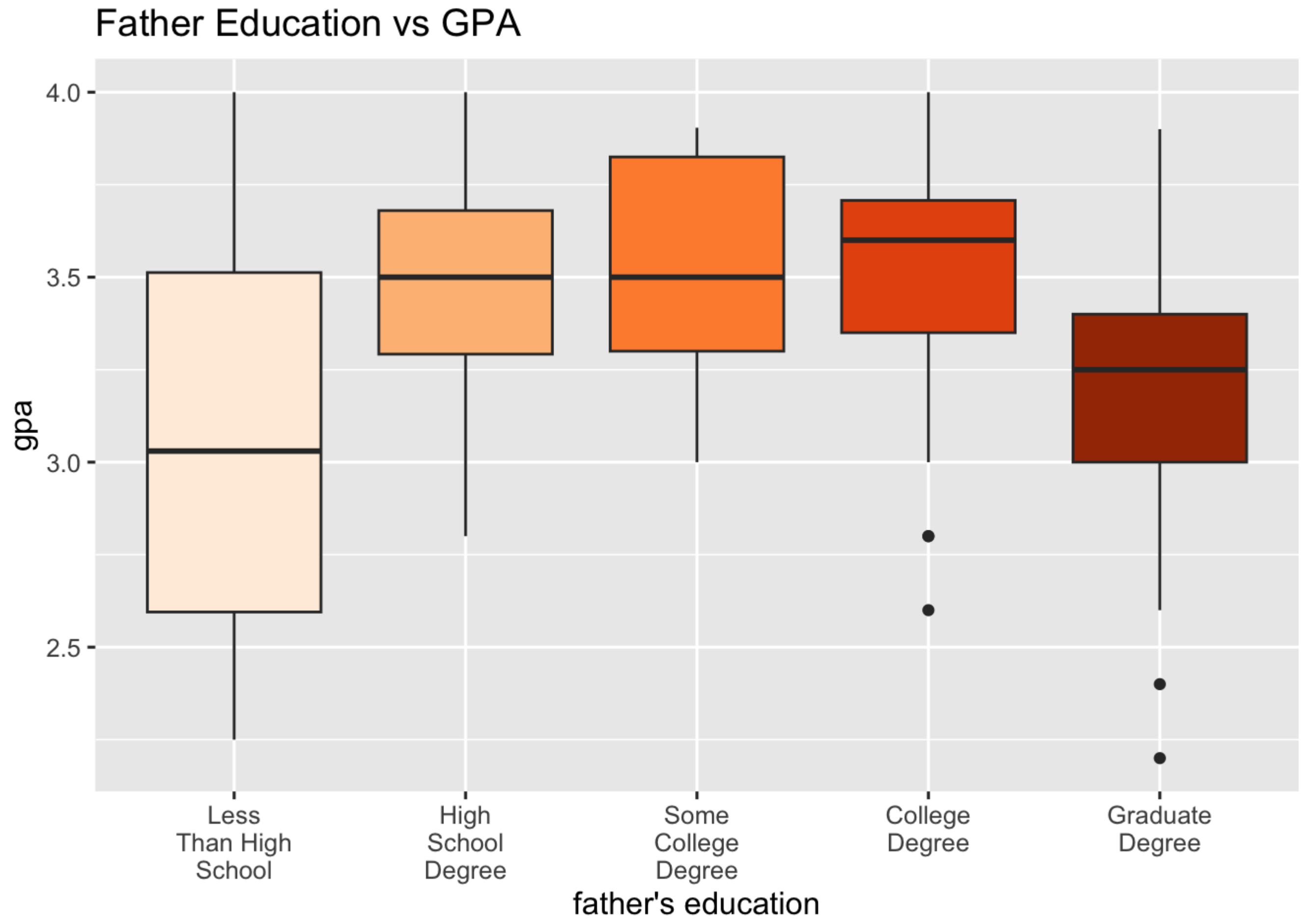
Analysis:

```r
food_cleaned |>

  #Reorder income
  mutate(
    income = fct_relevel(income, "< $15,000", "$15,001 to $30,000", "$30,001 to $50,000", "$50,001 to $70,000", "$70,001 to $100,000", "> $100,000")) |>

  #Instantiate ggplot
  ggplot() +

  #Specify x (income) and y (gpa). Fill plots with income
  aes(x = income, y=gpa, fill=income) +

  #Specify boxplot (legend is not needed)
  geom_boxplot(show.legend = FALSE) +


  #Fill categories using blue
  scale_fill_brewer(palette = "Blues") +

  #Wrap axis labels
  scale_x_discrete(labels = ~ str_wrap(as.character(.x), 10)) +

  #Set theme
  theme_grey(12) +

  #Set title
  ggtitle("Income vs GPA") +

  #Set axis labels
  labs(y = 'gpa', x = "income")
```

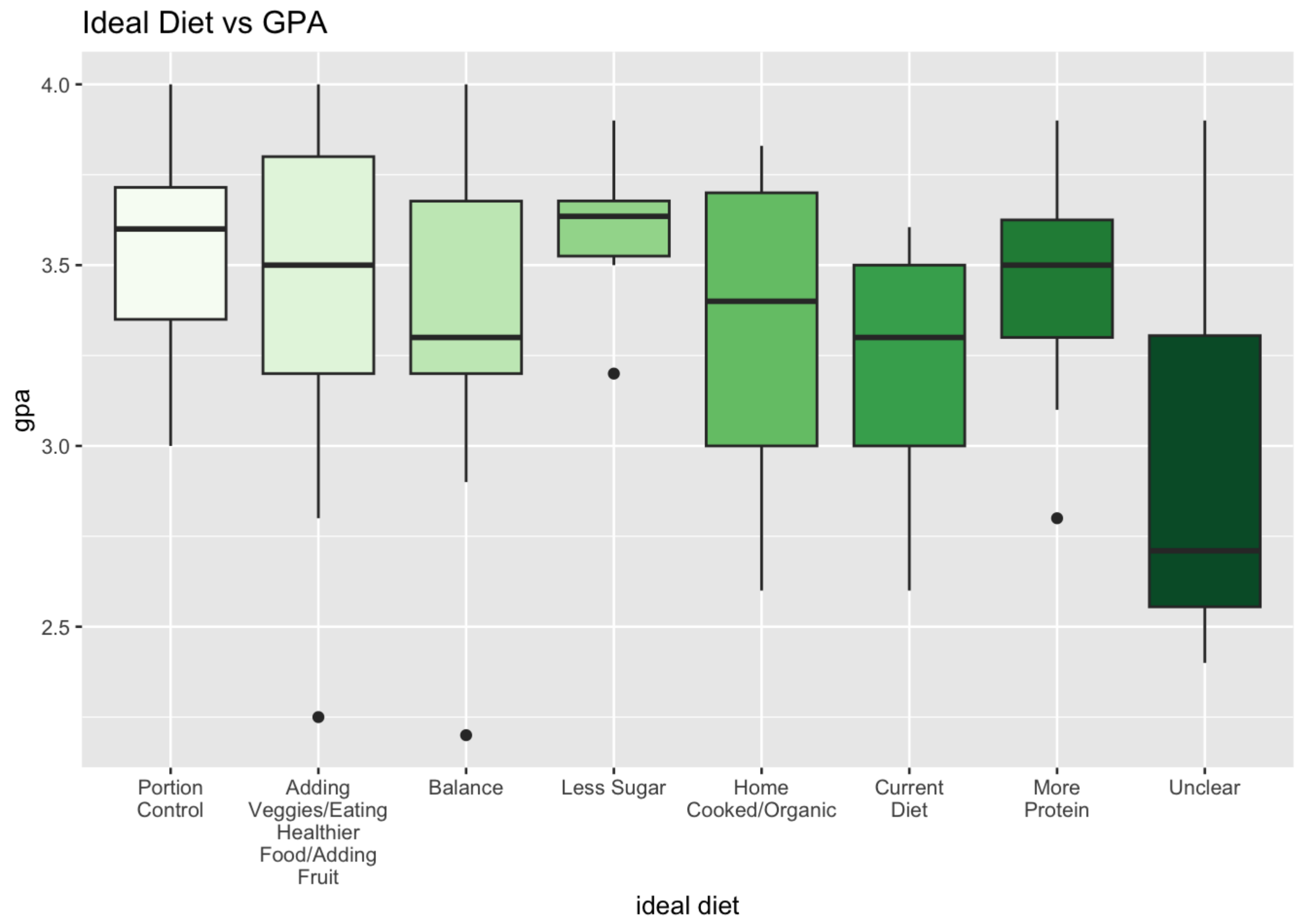


Income vs GPA

```r
food_cleaned |>

  #Reorder income
  mutate(
    father_education = fct_relevel(father_education, "Less Than High School", "High School Degree", "Some College Degree", "College Degree","Graduate Degree")) |>

  #Instantiate ggplot
  ggplot() +

  #Specify x (income) and y (gpa). Fill plots with income
  aes(x = father_education, y=gpa, fill=father_education) +

  #Specify boxplot (legend is not needed)
  geom_boxplot(show.legend = FALSE) +


  #Fill categories using blue
  scale_fill_brewer(palette = "Oranges") +

  #Wrap axis labels
  scale_x_discrete(labels = ~ str_wrap(as.character(.x), 10)) +

  #Set theme
  theme_grey(12) +

  #Set title
  ggtitle("Father Education vs GPA") +

  #Set axis labels
  labs(y = 'gpa', x = "father's education")
```


Father Education vs GPA
gpa
4.0
3.5
3.0
2.5
Less Than High School
High School Degree
Some College Degree
College Degree
Graduate Degree
father's education

```r
options(repr.plot.width = 10, repr.plot.height =3)

food_cleaned |>

  #Reorder income
  mutate(
    ideal_diet = fct_relevel(ideal_diet, "Portion Control", "Adding Veggies/Eating Healthier Food/Adding Fruit", "Balance", "Less Sugar", "Home Cooked/Organic", "Current Diet", "More Protein", "Unclear")) |>

  #Instantiate ggplot
  ggplot() +

  #Specify x (income) and y (gpa). Fill plots with income
  aes(x = ideal_diet, y=gpa, fill=ideal_diet) +

  #Specify boxplot (legend is not needed)
  geom_boxplot(show.legend = FALSE) +

  #Fill categories using blue
  scale_fill_brewer(palette = "Greens", labels = scales::label_wrap(10)) +

  #Wrap axis labels
  scale_x_discrete(labels = ~ str_wrap(as.character(.x), 10)) +

  #Set theme
  theme_grey(10) +

  #Set title
  ggtitle("Ideal Diet vs GPA") +


  #Set axis labels
  labs(y = 'gpa', x = "ideal diet")
```



Discussion:

For each of the explanatory variables, beginning with student income, average gpa appears to remain consistent regardless of student income level; however, both extremities (<$15,000 and >$100,000) appear to have the highest skewed distributions, where students in these categories are skewed towards having a lower GPA. This could be due to a few factors. One, there may be speculation that students in this cohort already come from low-income families and are not adequately provided with the resources needed to succeed in school growing up. Secondly, students in the >$100,000 cohort may lack the “drive” from doing well in school as they already have sufficient resources.

In regards to comparing father’s education and GPA, the pattern appears largely the same as student income level, but with greater differences between cohorts, which can likely be explained using the same reasons as above. A student who has a father with less than a high school degree may not be pushed enough to perform well in school; similarly, a student with a father who has at least a graduate degree may not have as high as a GPA as other cohorts due to a lack of “drive” from already having sufficient resources.

Lastly, as there is no inherent order for ideal diet, it is apparent that students who follow a diet fare better than students who don’t have a clear diet regimen.