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 × 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
## # ℹ 115 more rows
## # ℹ 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>, …
```

A detailed data dictionary for this dataset is available here. The dataset was originally downloaded from Kaggle, and you can find additional information about the dataset here.

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: *Your introduction here.* The modern student's life is a complex interplay of academic, personal, and socio-economic factors. Understanding how these elements interact can provide valuable insights into student behavior and performance. The dataset under examination comprises survey responses from students detailing their food choices, preferences, and personal information such as GPA, father's educational level, and income. Originating from a Kaggle source, this dataset is particularly intriguing due to its unclean nature, necessitating rigorous data cleaning prior to any analysis. For the purpose of this project, we will be focusing on the columns GPA, father_education, ideal_diet_coded, and income. The GPA column represents the academic performance of the student, father_education indicates the educational level of the student's father, ideal_diet_coded denotes the student's perception of what an ideal diet is, and income represents the student's income bracket. Together, these columns provide a multi-dimensional view of the student, potentially revealing patterns and correlations that might otherwise remain obscured.

Approach: *Your approach here.* Our primary objective is to discern if there's any correlation between a student's GPA and their income, the father's educational level, or the student's perception of what an ideal diet is. To achieve this, we will clean the dataset by handling missing values and outliers and use boxplots to visualize the distribution of GPA across different income brackets, father's educational levels, and ideal diet perceptions. Boxplots are chosen for their ability to succinctly represent the distribution of a continuous variable across different categories.

Analysis:

```
# Your R code here
# Data cleaning
food_cleaned <- food %>%
  filter(!is.na(GPA), !is.na(father_education), !is.na(ideal_diet_coded), !is.na(income)) %>%
  mutate(GPA = as.numeric(GPA),
         father_education = factor(father_education, levels = 1:5, labels = c("Less than HS", "HS Degree", "Some College", "College Degree", "Graduate Degree")),
         ideal_diet_coded = factor(ideal_diet_coded, levels = 1:8, labels = c("Portion Control", "Adding Veggies/Fruits", "Balance", "Less Sugar", "Home Cooked/Organic", "Current Diet", "More Protein", "Unclear")),
         income = factor(income, levels = 1:6, labels = c("<$15,000", "$15,001-$30,000", "$30,001-$50,000", "$50,001-$70,000", "$70,001-$100,000", ">$100,000")))
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `GPA = as.numeric(GPA)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Summary of columns
summary(food_cleaned$GPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.200   3.200   3.500   3.418   3.700   4.000       4
```

```
table(food_cleaned$father_education, useNA = "ifany")
```

```
##
##    Less than HS       HS Degree    Some College  College Degree Graduate Degree
##               4              34              12              46              27
```

```
table(food_cleaned$ideal_diet_coded, useNA = "ifany")
```

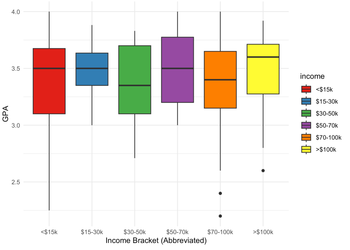
```
##
##       Portion Control Adding Veggies/Fruits                 Balance
##                    11                    44                      16
##            Less Sugar   Home Cooked/Organic           Current Diet
##                     6                    15                      13
##           More Protein                Unclear
##                    15                     3
```

```
table(food_cleaned$income, useNA = "ifany")
```

```
##
##         <$15,000  $15,001-$30,000  $30,001-$50,000  $50,001-$70,000
##                6                7               17               20
## $70,001-$100,000       >$100,000
##               32               41
```

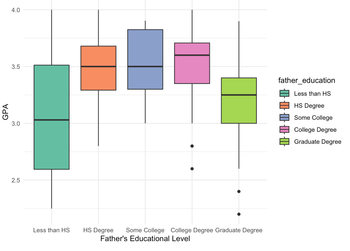
```
# Your R code here
# Ensure the income column is treated as a factor
food_cleaned$income <- as.factor(food_cleaned$income)

# Abbreviate the income levels
levels(food_cleaned$income) <- c("<$15k", "$15-30k", "$30-50k", "$50-70k", "$70-100k", ">$100k")

# Boxplot for GPA vs. abbreviated income with color
ggplot(food_cleaned, aes(x = income, y = GPA, fill = income)) +
  geom_boxplot() +
  xlab("Income Bracket (Abbreviated)") +
  ylab("GPA") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1")
```

```
## Warning: Removed 4 rows containing non-finite values (`stat_boxplot()`).
```



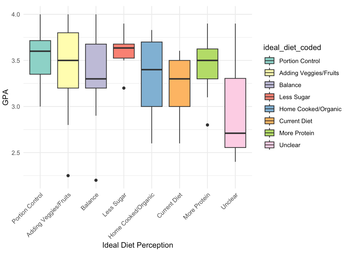

```
# Boxplot for GPA vs. father's educational level with color
ggplot(food_cleaned, aes(x = father_education, y = GPA, fill = father_education)) +
  geom_boxplot() +
  xlab("Father's Educational Level") +
  ylab("GPA") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2")
```

```
## Warning: Removed 4 rows containing non-finite values (`stat_boxplot()`).
```




```
# Boxplot for GPA vs. ideal diet perception with color
ggplot(food_cleaned, aes(x = ideal_diet_coded, y = GPA, fill = ideal_diet_coded)) +
  geom_boxplot() +
  xlab("Ideal Diet Perception") +
  ylab("GPA") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set3")
```

```
## Warning: Removed 4 rows containing non-finite values (`stat_boxplot()`).
```



Discussion: *Your discussion of results here.* Starting with the income bracket, it's evident that students from the higher income brackets, specifically those earning more than $100,000, tend to have a median GPA that is relatively high, around 3.6. This could suggest that students from wealthier backgrounds might have access to better educational resources or face fewer financial stresses, potentially leading to better academic performance. However, it's also worth noting that students from the lowest income bracket, less than $15,000, have a median GPA that is not significantly different from some higher income brackets. This could be indicative of the resilience and determination of students from lower-income backgrounds. When examining the father's educational level, there's a noticeable trend. Students whose fathers have a graduate degree tend to have a higher median GPA, around 3.5, compared to those whose fathers have less than a high school education, where the median GPA is closer to 3.0. This might suggest that parental education, and perhaps the value they place on academic achievement, could influence their children's academic performance. Lastly, when considering students' perception of an ideal diet, those who believe in "adding veggies/eating healthier food/adding fruit" and those who prefer "home cooked/organic" foods have a higher median GPA, both hovering around 3.5. This could imply a potential correlation between health-conscious choices and academic performance, though the causality is not clear. It's possible that students who prioritize their health also prioritize their studies, or that a healthier diet directly contributes to better cognitive function and academic performance. However, while these observations provide valuable insights, it's crucial to approach them with caution. The variability in GPA, as indicated by the spread of the boxes in the boxplots, shows that there's a wide range of GPAs within each category. For instance, in the higher income bracket, while the median GPA is around 3.6, there are students with GPAs as low as 2.8 and as high as 4.0. This variability underscores the fact that while there might be trends, individual experiences and performances can vary widely. In conclusion, while there are apparent relationships between GPA and the three variables of interest, it's essential to consider the broader context and other potential influencing factors. Furthermore, outliers in some income brackets might indicate exceptional cases where students either significantly outperform or underperform compared to their peers in the same income group. It's also worth noting that the data is self-reported, which might introduce biases or inaccuracies.