Project 3

This is the dataset you will be working with:

```
food <- readr::read_csv("https://wilkelab.org/DSC385/datasets/food_coded.csv",
  na = c("","Personal","Unknown","NA"))
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>, ...
```

```
food["ideal_diet_coded"]
```

```
## # A tibble: 125 x 1
##    ideal_diet_coded
##               <dbl>
##  1                8
##  2                3
##  3                6
##  4                2
##  5                2
##  6                2
##  7                2
##  8                2
##  9                6
## 10                2
## # ... with 115 more rows
```

```
summary(food$ideal_diet_coded)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.704   6.000   8.000
```

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.

Student's Income Level

#1 - less than $15,000 #2 - $15,001 to $30,000 #3 - $30,001 to $50,000 #4 - $50,001 to $70,000 #5 - $70,001 to $100,000 #6 - higher than $100,000

#Father's Education #1 - less than high school #2 - high school degree #3 - some college degree #4 - college degree #5 - graduate degree

#Ideal Diet Coded #1 – portion control #2 – adding veggies/eating healthier food/adding fruit #3 – balance #4 – less sugar #5 – home cooked/organic #6 – current diet #7 – more protein #8 – unclear

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: *We are working with the `Food Coded.csv` dataset, which consists of food choices and preferences of 127 college students. In this dataset, each record records grades and food preferences of student. There are 60 columns that provides information about GPA, gender, calorie-info, breakfast preference, nutrition, association with foods, income, education and profession of parents. Some column-values are encoded as integers. For this project, we will explore whether there is a relationship between GPA and student income, the father's educational level, or the student's perception of what an ideal diet is. To carry out this objective, we will make use of 4 columns from the dataset. First column is `GPA` which stores the GPA in numeric format, Second is `income`,encoded in the range of 1 to 6. Third column is `father_education`, encoded in the range of 1 to 5. The fourth column is `ideal_diet_coded`, encoded in the range of 1 to 8. First, we will clean the dataset and transform it into easily readable format. Then, we will plot a jittered box-plot for GPA vs income, GPA vs father_education and GPA vs ideal_diet_coded to understand whether they are related or not*

Approach: *To clean the data, first we will use `select()` function to select the four relevant columns. GPA column is in character format in the original dataset. So, we will change the GPA to numeric format by using `as.numeric()`. As discussed in the introduction, 3 relevant columns (ie. `income`, `father_education`,* *`ideal_diet_coded`) are encoded in the integer format. So, to convert the data of these 3 column into categorical variable, we will mutate the income, father_education and ideal_diet_coded column using `case_when`. Moreover, the categories in `income` and `father_education` has order to it. Therefore, we will arrange it in an reasonable order using `fct_relevel()` method.*

Since the objective is to plot the relationship between a categorical variable and a quantitative variable, jittered side-by-side boxplots would be a great choice. Reasoning is that it's very handy to compare groups (`income`, `father_education`, `ideal_diet_coded`) on a numerical variable(`GPA`) using side-by-side boxplots. Adding jitter points on it makes it very easy and intuitive to visualize the data.

Analysis:

```
# Data Cleaning Code

food_cleaned <- food %>%
  # select the 4 relevant columns
  select(GPA, father_education, income, ideal_diet_coded) %>%
  # convert GPA to numeric
  mutate(GPA = as.numeric(GPA))%>%
  mutate(  # decode income variable to meaningful categories
    income = 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_
      )
    )%>%
  # put the income into ordered categories
  mutate(income = fct_relevel(income,
                          "less than $15,000",
                          "$15,001 to $30,000",
                          "$30,001 to $50,000",
                          "$50,001 to $70,000",
                          "$70,001 to $100,000",
                          "higher than $100,000"))%>%
  # decode father_education variable to meaningful categories
  mutate(
    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",
     TRUE ~ NA_character_
      )
    )%>%
    # put the father_Education into ordered categories
  mutate(father_education = fct_relevel(father_education,
                          "less than high school",
                          "high school degree",
                          "some college degree",
                          "college degree",
                          "graduate degree"
                          ))%>%
  mutate( #decode ideal_diet_coded variable to meaningful categories
    ideal_diet_coded = 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",
     TRUE ~ NA_character_
    )
  )
```

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

```
# Summary Function for numerical variable.
summary(food_cleaned$GPA)
```

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

```
# Table for Categorical Variables
table(food_cleaned$income, useNA = "ifany")
```

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

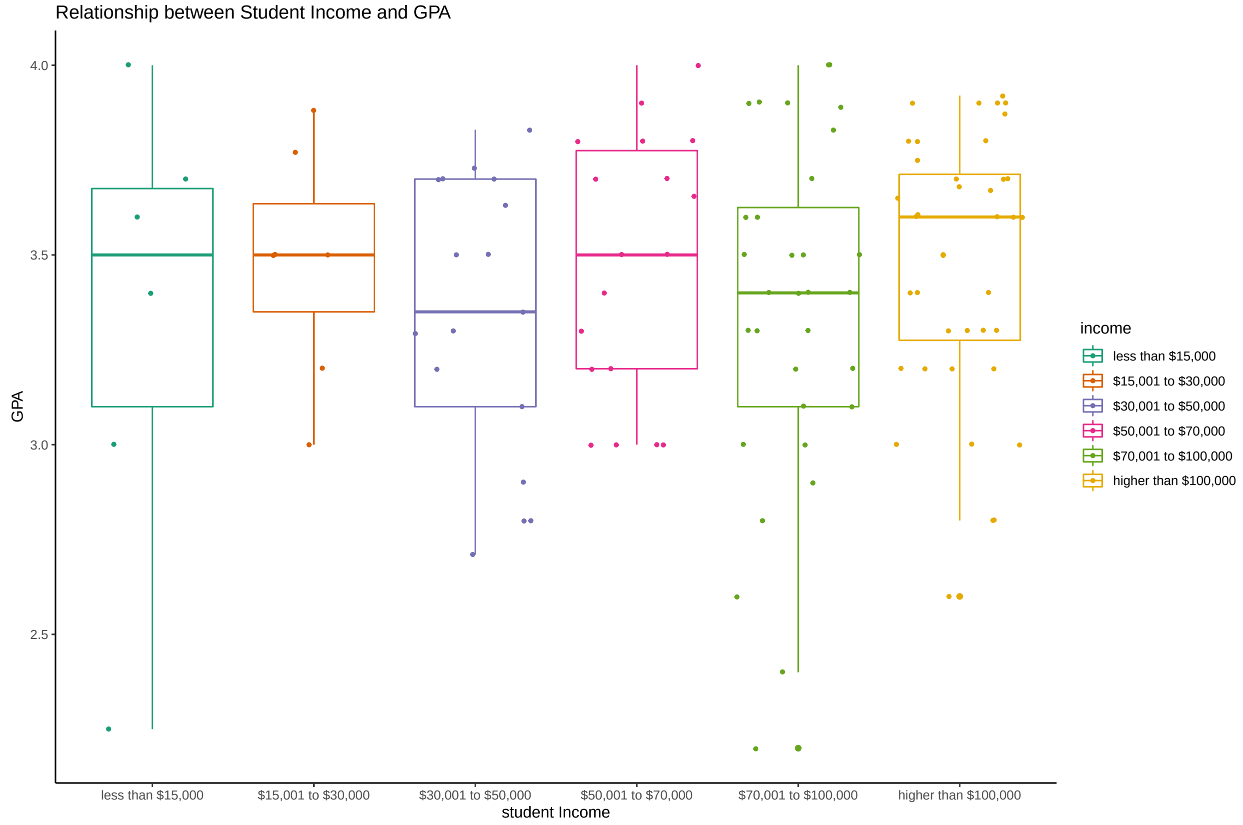
```
table(food_cleaned$father_education, useNA = "ifany")
```

```
##
## less than high school    high school degree   some college degree
##                     4                    34                    12
##        college degree       graduate degree                  <NA>
##                    46                    28                     1
```

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

```
##
## adding veggies/eating healthier food/adding fruit
##                                                44
##                                           balance
##                                                17
##                                      current diet
##                                                13
##                               home cooked/organic
##                                                15
##                                        less sugar
##                                                 6
##                                      more protein
##                                                    16
##                                       portion control
##                                                    11
##                                              unclear
##                                                     3
```

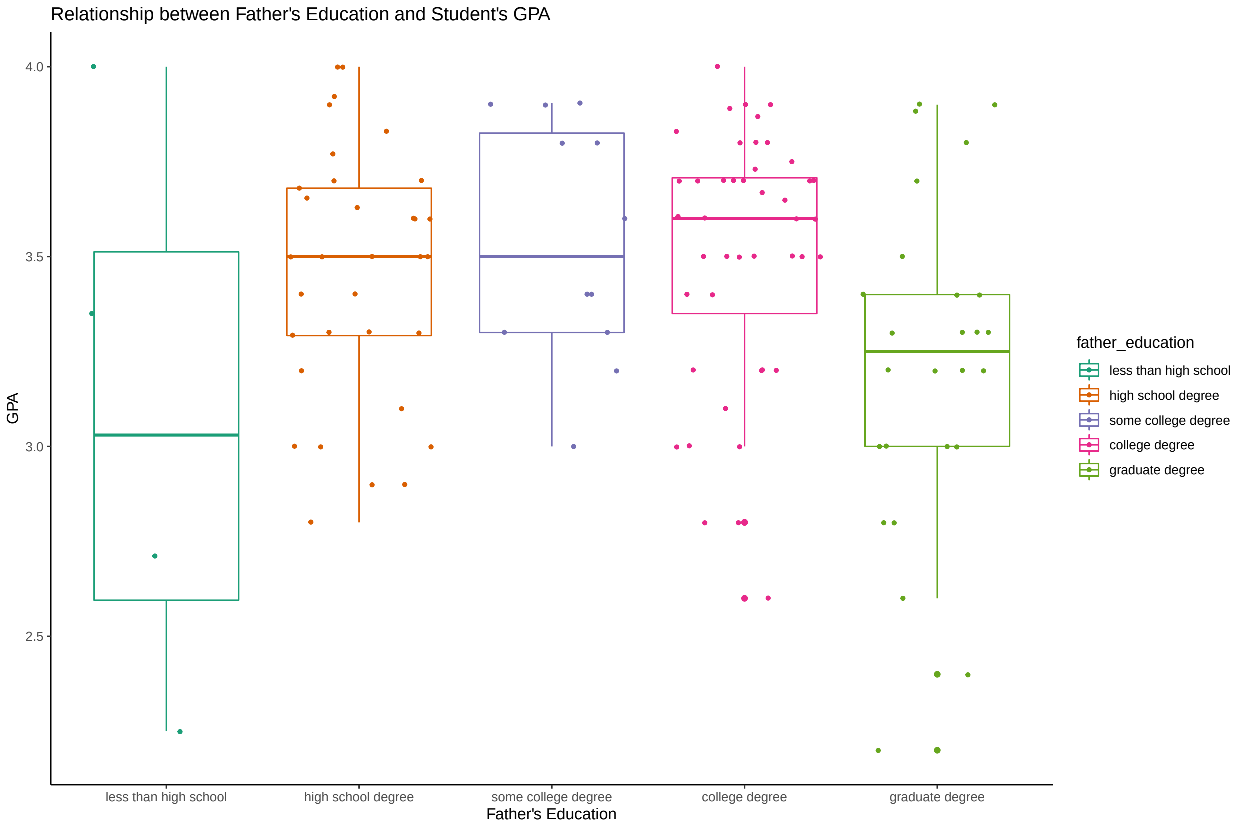
```
# Jittered Boxplot that shows the relationship between GPA and
# Student's Income
ggplot(data = subset(food_cleaned, !is.na(GPA), !is.na(income)), aes(income, GPA, colour=income)) +
  geom_boxplot()+
  geom_point(position = "jitter", size = 1)+
  scale_color_brewer(palette="Dark2")+
  labs(title="Relationship between Student Income and GPA",x="student Income", y = "GPA")+
  theme_classic()
```



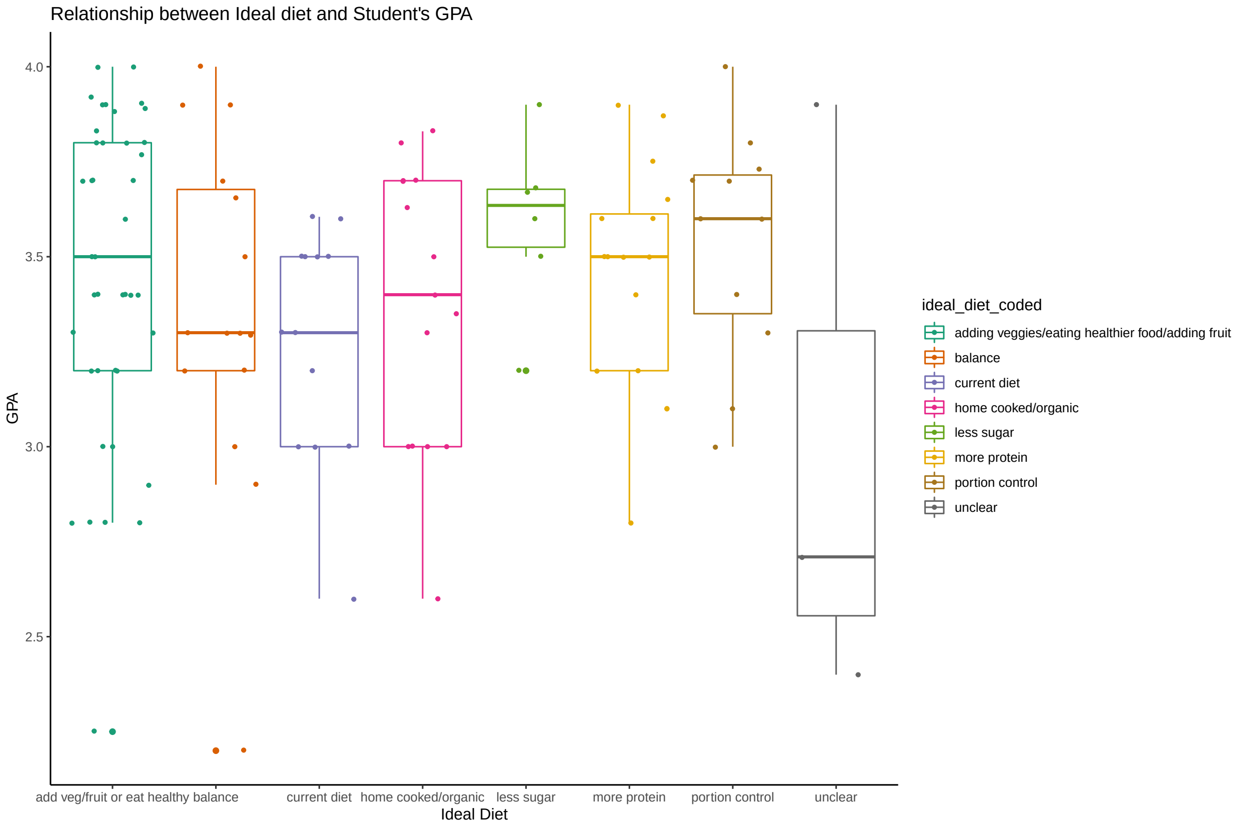
```
# Jittered Boxplot that shows the relationship between Student's GPA and
# Father's Education
ggplot(data = subset(food_cleaned, !is.na(father_education), !is.na(GPA)), aes(father_education, GPA, c
  geom_boxplot()+
  geom_point(position = "jitter", size = 1)+
  scale_color_brewer(palette="Dark2")+
  labs(title="Relationship between Father's Education and Student's GPA",x="Father's Education", y = "G
  theme_classic()
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).

## Warning: Removed 5 rows containing missing values (geom_point).
```



```
# Jittered Boxplot that shows the relationship between Student's GPA and
# Ideal diet choice for students
ggplot(data = subset(food_cleaned, !is.na(GPA), !is.na(ideal_diet_coded)), aes(ideal_diet_coded, GPA, c
  geom_boxplot()+
  geom_point(position = "jitter", size = 1)+
  scale_x_discrete(labels=c("adding veggies/eating healthier food/adding fruit" = "add veg/fruit or eat
  scale_color_brewer(palette="Dark2")+
  labs(title="Relationship between Ideal diet and Student's GPA",x="Ideal Diet", y = "GPA")+
  theme_classic()
```



Discussion: *For GPA vs student income, it's highly unlikely to say that GPA and student income are related, since the mean of all the groups are between 3.3 and 3.8. There is no huge noticeable variance when it comes to mean. But outliers are spotted for the income group less than $15,000, between $70,001 to $100,000 and higher than $100,000. Out of this, outliers in less than $15,000 are seen on both extreme ends but overall very few points are seen for that category. There are more people for categories who has income greater than $30,000 and for these categories mean is nearly same or close.*

For GPA vs father's education, one can see that for father's education less than high school, there are less students and mean GPA is also less(ie. nearly 3). Whereas in rest other categories, there are relatively more students and GPA is nearly same.(around 3.5)

For GPA vs ideal diet preference, there are more students who have opted for adding veggies category and the mean is around 3.5. Mean in other categories is nearly same or close to 3.5 except for the last category unclear. For unclear category, mean is very low (ie. below 3.0)

To conclude, one can say that for categories having income greater than $30,000, GPA is nearly close, for father's education greater than high school, GPA is close and when it comes to ideal diet choice, except the category of unclear, GPA is nearly close. No concrete conclusion can be made for other categories since the no. of people participating in these categories are very less. So, it can be speculated that when it comes to students of the same college, generally father's education is greater than high school, and income level of student is greater than $30,000