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("", "NA", "#N/A", "Personal","Unknown", "nan", "NaN"))

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           NA            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           NA            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 <dbl>, cook <dbl>,
## #   comfort_food_reasons_coded_1 <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>, fries <dbl>,
## #   fruit_day <dbl>, grade_level <dbl>, greek_food <dbl>,
## #   healthy_feeling <dbl>, healthy_meal <chr>, ideal_diet <chr>,
## #   ideal_diet_coded <dbl>, income <dbl>, indian_food <dbl>,
## #   italian_food <dbl>, life_rewarding <dbl>, marital_status <dbl>,
## #   meals_dinner_friend <chr>, mother_education <dbl>, mother_profession <chr>,
## #   nutritional_check <dbl>, on_off_campus <dbl>, parents_cook <dbl>,
## #   pay_meal_out <dbl>, persian_food <dbl>, self_perception_weight <dbl>,
## #   soup <dbl>, sports <dbl>, thai_food <dbl>, tortilla_calories <dbl>,
## #   turkey_calories <dbl>, type_sports <chr>, veggies_day <dbl>,
## #   vitamins <dbl>, waffle_calories <dbl>, weight <chr>
```

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?

Introduction: The dataset we are looking at today is called Food. It is a survey taken from a University. This dataset includes information on food choices, nutrition, preferences, childhood favorites, and other information from college students. The question we are interested in answering with this dataset is “Is GPA related to student income, the father’s educational level, or the student’s perception of what an ideal diet is?”.

This survey originally asks 125 students 61 different questions, thus 61 different varaibles. We do not need all 61 to tackle our question at hand. To answer this we only need: “GPA”, “Income”, “Father_education”, and “ideal_diet_coded”. GPA gives us information about the student’s GPA (between 0 and 4), Income gives us data regarding the student’s family income as a categorical variable (from less than $15,000 to greater than $100,000), Father_education gives us data about the father of the student’s highest level of education as a categorical variable (from less than HS to graduate degree), and finally Ideal_diet_coded details the students ideal diet as a categorical variable (from more protein to more veggies to what they are currently eating, etc.).

Approach: To start our dive into this question, we first have to look at the data. We begin by using select to choose the relevant variables we need. After obtaining our new dataset, we have to clean it. To clean it, we make sure all of our missing variables are accounted for and make sure GPA is coded as a numeric variable and the other three are coded as categorical variables. Once we have a clean dataset and all of our variables look as we want them, we can start to answer our question at hand.

To tackle this question, we want to make 3 different ridgeline plots, comparing GPA vs. income, GPA vs. father’s education, and GPA vs. ideal diet. We want to make ridgeline plots because these we have one numerical variable plotted against a categorical variable. Using a ridgeline plot will allow us to see if there is a relationship between the variables by looking at the distribution among each group, thus answering our question.

Analysis:

```
# new dataset with only 4 relevant variables
food1 <- food %>%
  select(GPA, income,  father_education,  ideal_diet_coded)

# cleaning GPA to be read as a numerical
food1$GPA[74]=3.79
food1$GPA <- as.double(food1$GPA)

# getting income, fathers education, and ideal diet to be read as categorical
food1 <- food1 %>%
  mutate(
    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_ # should never reach
    ),
    father_education = case_when(
      father_education == 1 ~ "Less Than HS",
      father_education == 2 ~ "HS Degree",
      father_education == 3 ~ "Some College/Assoc. Degree",
      father_education == 4 ~ "Bachelors",
      father_education == 5 ~ "Graduate Degree",
      TRUE ~ NA_character_ # should never reach
    ),
    ideal_diet_coded = case_when(
      ideal_diet_coded == 1 ~ "Portion Control",
      ideal_diet_coded == 2 ~ "Veggies/Fruits/Healthy",
      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_ # should never reach
    ),
    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"),
    father_education = fct_relevel(father_education, "Less Than HS", "HS Degree", "Some College/
Assoc. Degree", "Bachelors", "Graduate Degree"),
    ideal_diet_coded = fct_rev(fct_infreq(ideal_diet_coded))
  )

summary(food1$GPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.200   3.200   3.500   3.419   3.700   4.000       4
```

```
table(food1$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(food1$father_education, useNA = "ifany")
```

```
##
##               Less Than HS                  HS Degree
##                          4                         34
## Some College/Assoc. Degree                  Bachelors
##                         12                         46
##            Graduate Degree                       <NA>
##                         28                          1
```

```
table(food1$ideal_diet_coded, useNA = "ifany")
```

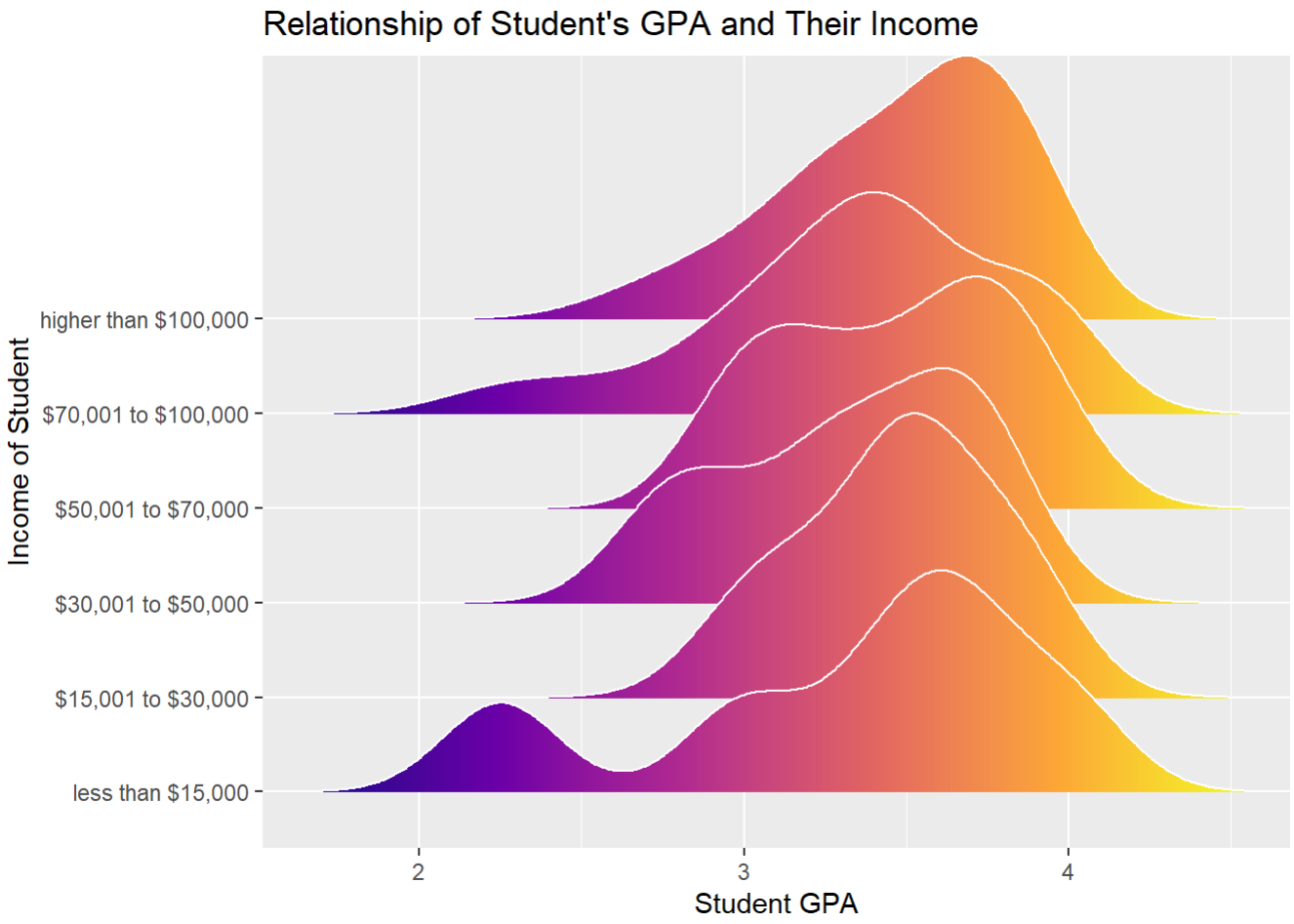
```
##
##                Unclear             Less Sugar        Portion Control
##                      3                      6                     11
##           Current Diet    Home Cooked/Organic           More Protein
##                     13                     15                     16
##                Balance Veggies/Fruits/Healthy
##                     17                     44
```

GPA has 4 missing responses, income had 1 missing, father's education had 1 missing and diet had none missing. Average GPA was 3.5, with a skewed left distribution, a couple gaps worth looking more into. The counts for income increased with each level increase, highest counts were higher than $100,000 group and lowest count was less than $15,000 group. For education level of father, few had less than HS so that could lead to some interesting numbers, and most had dads with HS degree or Bachelors. For diet, majority thought adding veggies, fruits, or eating healthier was the best, however, all groups except for less sugar and unclear had less than 10 counts. For ordering, I had income be in order of income, education level be in order of increasing education level, and for diet was ordered as whowever answered that question more.

These tables showed us that since groups between each variable have very different counts, this might lead to us not being able to see a true relationship, for example, for the diet question, only 3 answered unclear, while 44 said veggies/fruits/healthy. This will for sure skew our analysis and the plots. To get more accurate results, would like to see more, even counts between groups.

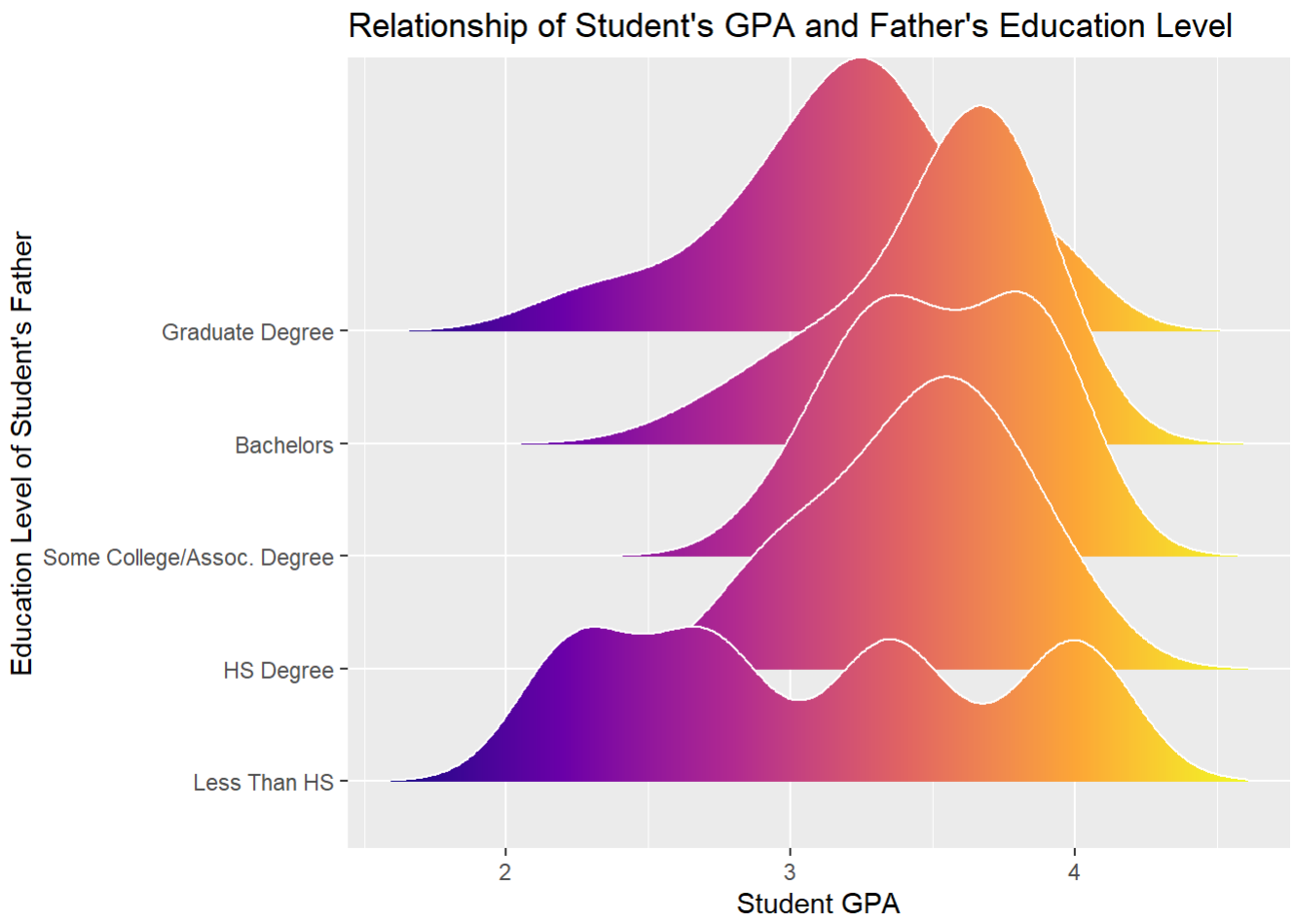
```
ggplot(data = subset(food1, !is.na(GPA)), aes(x = GPA, y = income, fill = stat(x))) +
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.001, color = "white") +
  scale_fill_viridis_c(name = "GPA", option = "C") +
  xlab("Student GPA") +
  ylab("Income of Student")+
  ggtitle("Relationship of Student's GPA and Their Income")+
  theme(legend.position="none")
```

```
## Picking joint bandwidth of 0.18
```



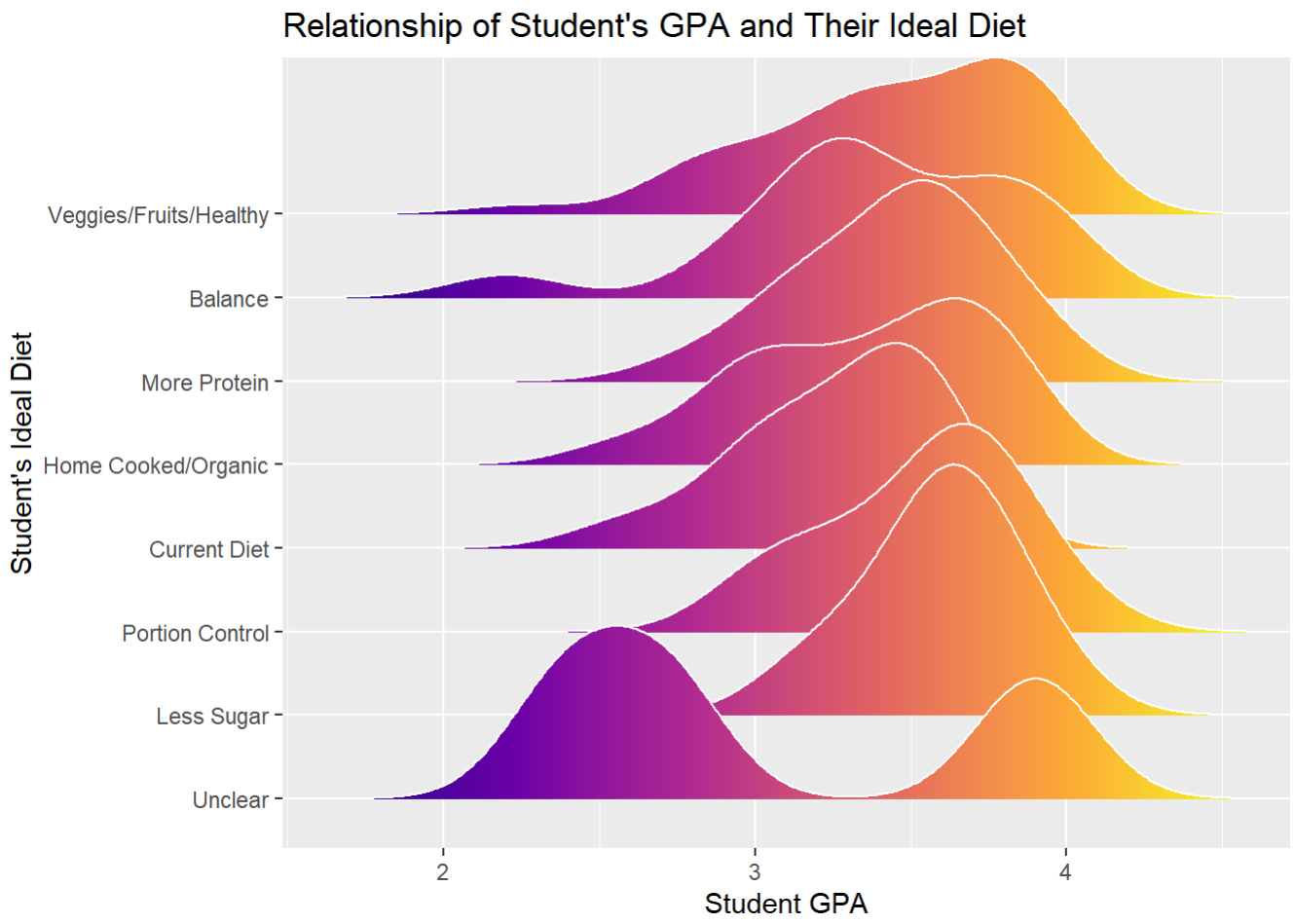

```
ggplot(data = subset(food1, !is.na(father_education), !is.na(GPA)), aes(x = GPA, y = father_educ
ation, fill = stat(x))) +
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.001, color = "white") +
  scale_fill_viridis_c(name = "GPA", option = "C") +
  xlab("Student GPA") +
  ylab("Education Level of Student's Father")+
  ggtitle("Relationship of Student's GPA and Father's Education Level")+
  theme(legend.position="none")
```

```
## Picking joint bandwidth of 0.203
```




```
ggplot(data = subset(food1, !is.na(GPA)), aes(x = GPA, y = ideal_diet_coded, fill = stat(x))) +
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.001, color = "white") +
  scale_fill_viridis_c(name = "GPA", option = "C") +
  xlab("Student GPA") +
  ylab("Student's Ideal Diet") +
  ggtitle("Relationship of Student's GPA and Their Ideal Diet")+
  theme(legend.position="none")
```

```
## Picking joint bandwidth of 0.19
```



Discussion: For the analysis, I decided to take out missing cells because GPA was being compared to the other variables and I thought that 4 out of 125 was not that many and it would skew the data where the variables would be unbalanced for each ridgeline plot.

First, we'll look at the relationship between GPA and Income based off the ridgeline plot. There does not seem to be any type of relationship between these two variables because they all had their peak of the distribution around the 3.5 GPA. The distribution of GPA for each income level was about the same, however, the "Less than $15,000" group (which was bimodal) had an interesting second peak at the 2.25 GPA which was the lowest peak among any group. Hard to say if making less money meant a lower GPA because plenty of people in the same group had a similar GPA to the other groups.

Next, the relationship between GPA and the Education Level of the Father. Again looking at the plot, there does not seem to be any relationship between the two variables. It might seem like a father with some type of Graduate level degree had a student with a little bit lower GPA, like .25 points lower, but more analysis would have to be done to see if that is significant. A father with no HS degree had students with GPA's all over the place, low to high. The "HS degree", "Some college/Assoc Degree", and "Bachelors" group had very similar distributions and peaks, not much variance to those 3 groups but the other 2 like previously stated, had a wide variance.

Lastly, for the relationship between GPA and the Student's Ideal Diet, we look at the plot. There, again, does not appear to be any relationship between the ideal diets of students and their GPA, as every group except for the unclear group, has similar distribution and peaks. Would not be able to tell GPA based off a student's ideal diet. Again, we see for unclea the two peaks because only 3 people answered as unclear and for veggies/fruits/healthy 44 answered, so that could be why we see the distribution to be more spread.