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 calor…¹ calor…² calor…³ coffee comfo…⁴ comfo…⁵ comfo…⁶
##    <chr>  <dbl>     <dbl>   <dbl>   <dbl>   <dbl>  <dbl> <chr>   <chr>     <dbl>
##  1 2.4        2         1     430     NaN     315      1 none    we don…       9
##  2 3.654      1         1     610       3     420      2 chocol… Stress…       1
##  3 3.3        1         1     720       4     420      2 frozen… stress…       1
##  4 3.2        1         1     430       3     420      2 Pizza,… Boredom       2
##  5 3.5        1         1     720       2     420      2 Ice cr… Stress…       1
##  6 2.25       1         1     610       3     980      2 Candy,… None, …       4
##  7 3.8        2         1     610       3     420      2 Chocol… stress…       1
##  8 3.3        1         1     720       3     420      1 Ice cr… I eat …       1
##  9 3.3        1         1     430     NaN     420      1 Donuts… Boredom       2
## 10 3.3        1         1     430       3     315      2 Mac an… Stress…       1
## # … with 115 more rows, 51 more variables: 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>, father_profession <chr>, fav_cuisine <chr>,
## #   fav_cuisine_coded <dbl>, fav_food <dbl>, food_childhood <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?

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: *Food is a dataset of responses to a survey on food preferences from 126 college students at Mercyhurst University. In the survey students were asked about their GPA, food preferences, their idea of a good diet, their current diet, as well as demographic information such as the education level of their parents, their personal income, their gender, etc. This analysis plans to answer the following question: Is GPA related to student income, the father's educational level, or the student's perception of what an ideal diet is? We do this by analyzing the following columns from food: GPA (from 0.0 to 4.0, continuous), income (categorically binned in increments of $15,000), father_education (categorical) and ideal_diet_coded (categorical, with values such as "less sugar", "more protein" and "portion control").*

Approach: *The approach is pretty straightforward, first we'll make a boxplot of GPA for each group. Then we need to do some hypothesis testing to see differences in means between groups. With the way the analysis is set up, it's screaming for us to run a one-way ANOVA on the data, so we'll do that. Responses are independent, so that is satisfied and we'll need to do some analysis for common variance and normally distributed. ANOVA was chosen over using notched boxplots because they came off the hinges when rendered causing ugly visualizations, and it's also just a more exhaustive method of determining differences between groups, based on a quantitative response variable*

Analysis:

Below is a breakdown of each of the four columns after data cleaning

```
data <-food %>% transmute(
  GPA=GPA,
  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'),
  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_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')
) %>% transmute(
  GPA = as.numeric(GPA),
  income = as.factor(income),
  father_education = as.factor(father_education),
  ideal_diet_coded = as.factor(ideal_diet_coded)
) %>% mutate(
  father_education = fct_relevel(father_education, "Less than high school", "High school
degree", "Some college degree", "College degree", "Graduate degree"),
  income = fct_relevel(income, "Less than $15,000", "$15,001 to $30,000","$30,001 to $5
0,000","$50,001 to $70,000","$70,001 to $100,000","Higher than $100,000")
)
```

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

```
summary(data$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(data$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
```

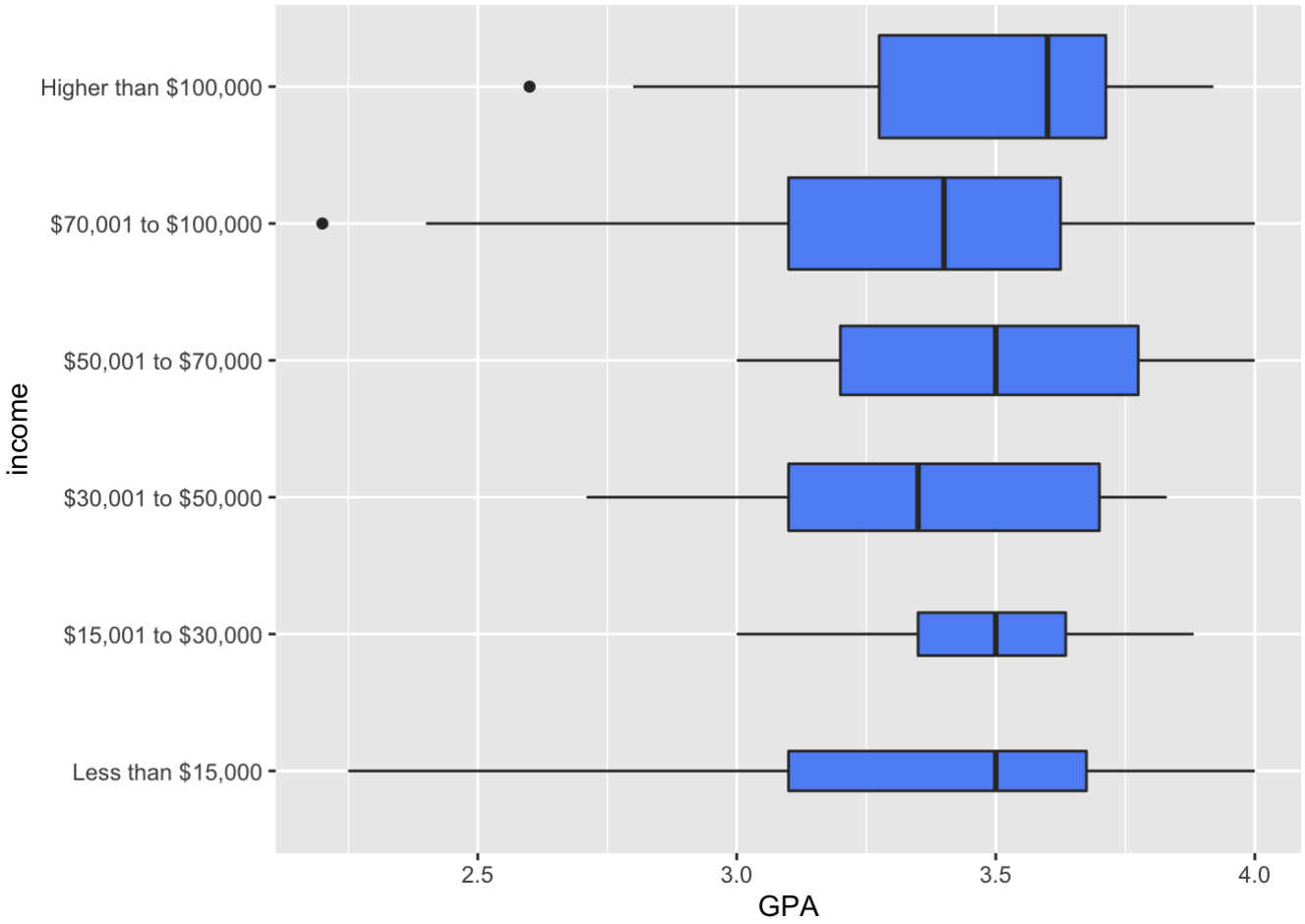
```
table(data$father_education, useNA='ifany')
```

```
##
## Less than high school    High school degree   Some college degree
##                     4                    34                    12
##        College degree       Graduate degree                  <NA>
##                    46                    28                     1
```

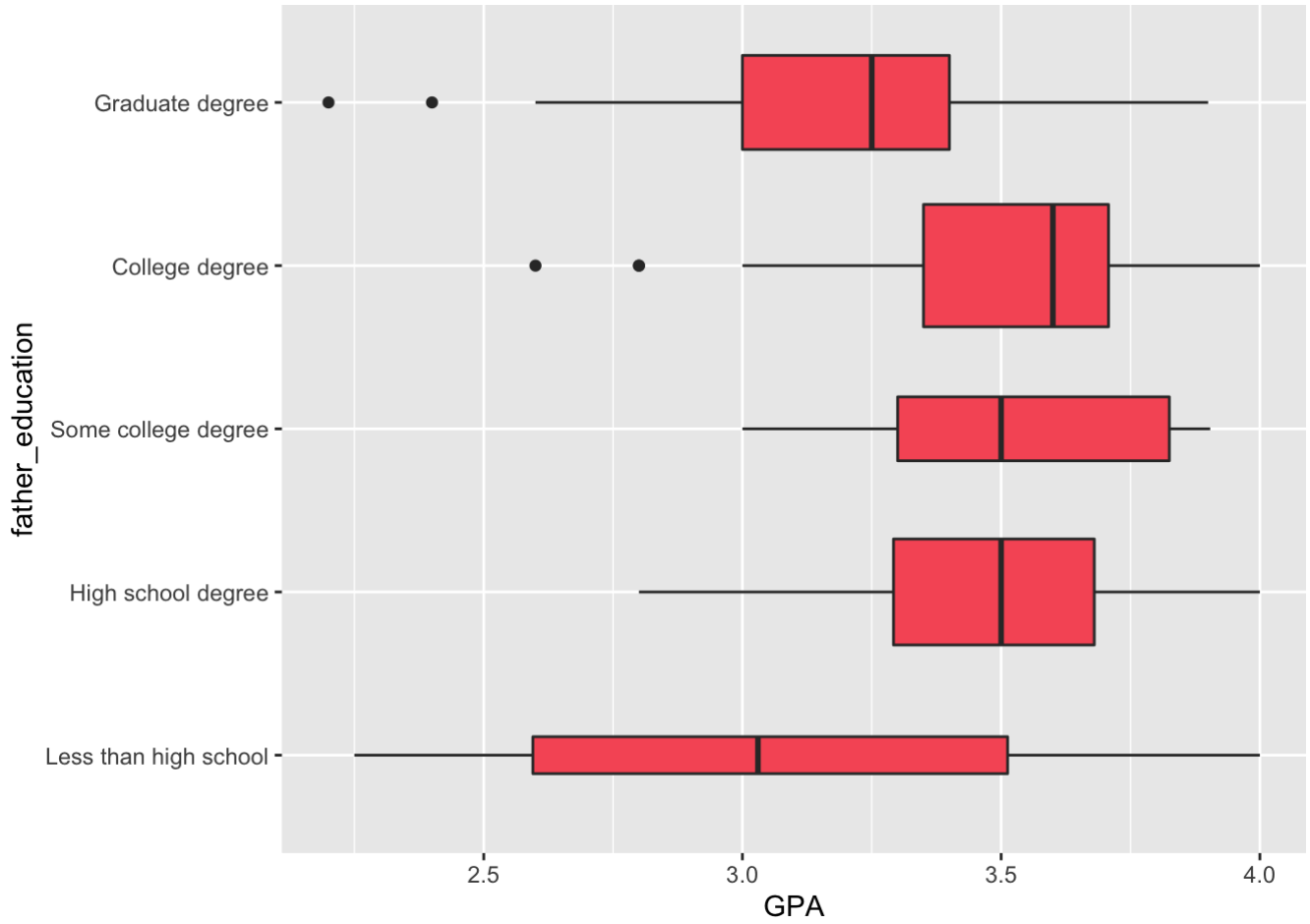
```
table(data$ideal_diet_coded, useNA='ifany')
```

```
##
## 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
```

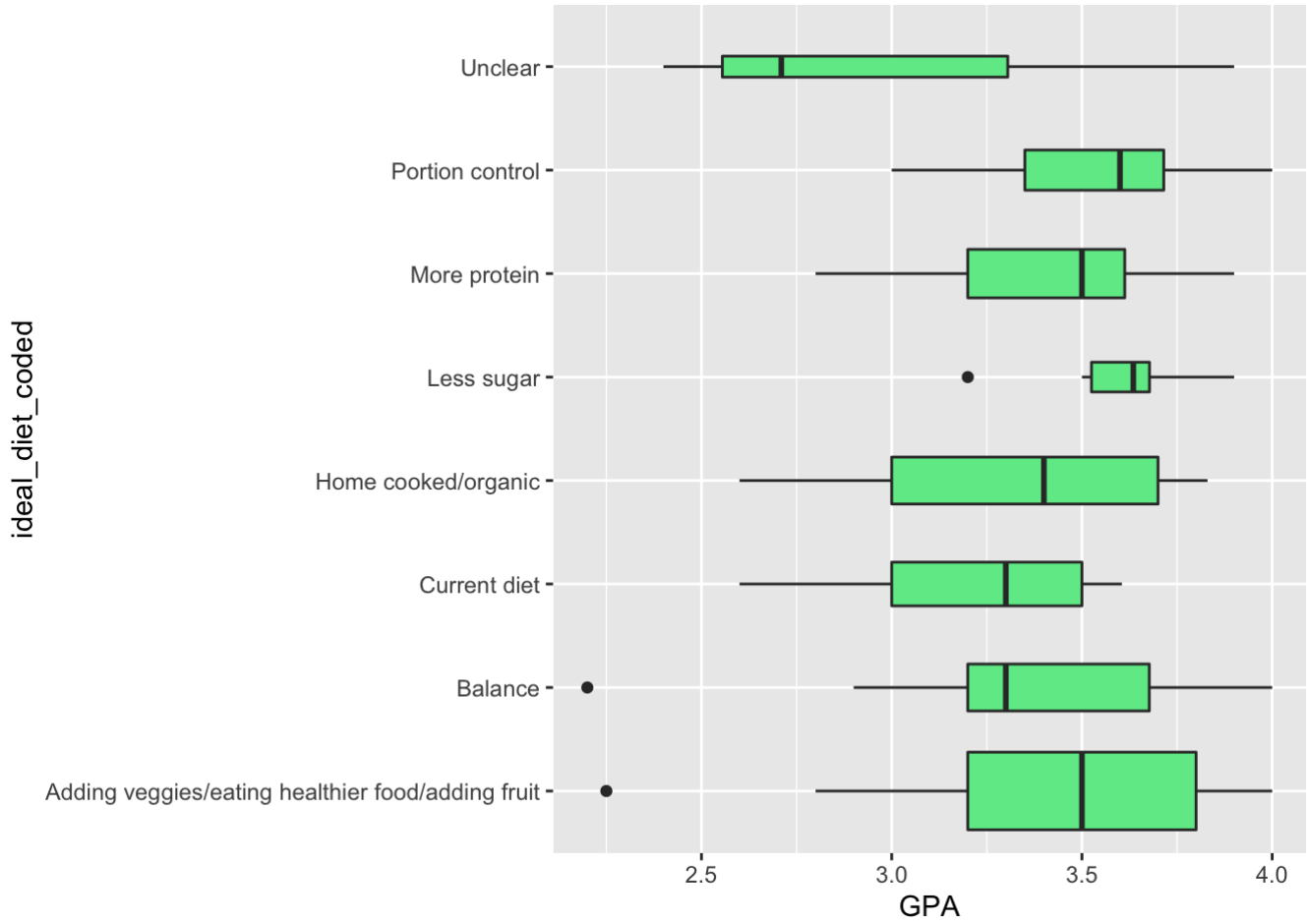
```
data %>% drop_na(income) %>% drop_na(GPA) %>%
  ggplot(aes(y=income,x=GPA)) + geom_boxplot(varwidth = TRUE, fill='#6092F5') + theme(le
gend.position="none")
```




```
data %>% drop_na(father_education) %>% drop_na(GPA) %>%
  ggplot(aes(y=father_education,x=GPA)) + geom_boxplot(varwidth = TRUE, fill='#F75B65')
 + theme(legend.position="none")
```




```
data %>% drop_na(ideal_diet_coded) %>% drop_na(GPA) %>%
  ggplot(aes(y=ideal_diet_coded,x=GPA)) + geom_boxplot(varwidth = TRUE, fill='#6DE996')
 + theme(legend.position="none")
```




```
one_way_income <- aov(GPA ~ income, data = data)
one_way_father_edu <- aov(GPA ~ father_education, data = data)
one_way_diet <- aov(GPA ~ ideal_diet_coded, data = data)

summary(one_way_income)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## income        5  0.501  0.1003   0.649  0.663
## Residuals   114 17.611  0.1545
## 5 observations deleted due to missingness
```

```
summary(one_way_father_edu)
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## father_education   4  2.071  0.5178   3.703 0.00715 **
## Residuals        114 15.941  0.1398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 6 observations deleted due to missingness
```

```
summary(one_way_diet)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## ideal_diet_coded   7  1.288  0.1840   1.225  0.295
## Residuals        112 16.825  0.1502
## 5 observations deleted due to missingness
```

```
leveneTest(GPA ~ income, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   5   0.823 0.5358
##       114
```

```
aov_residuals <- residuals(object = one_way_income)
shapiro.test(x = aov_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.96519, p-value = 0.003382
```

```
leveneTest(GPA ~ father_education, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   4  2.2714 0.06583 .
##       114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov_residuals <- residuals(object = one_way_father_edu)
shapiro.test(x = aov_residuals )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.98085, p-value = 0.08806
```

```
leveneTest(GPA ~ ideal_diet_coded, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   7  1.2327 0.2909
##       112
```

```
aov_residuals <- residuals(object = one_way_diet)
shapiro.test(x = aov_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.97055, p-value = 0.009805
```

```
kruskal_income = kruskal.test(GPA ~ income, data = data)
kruskal_income
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  GPA by income
## Kruskal-Wallis chi-squared = 2.3473, df = 5, p-value = 0.7993
```

```
kruskal_diet = kruskal.test(GPA ~ ideal_diet_coded, data = data)
kruskal_diet
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  GPA by ideal_diet_coded
## Kruskal-Wallis chi-squared = 6.8299, df = 7, p-value = 0.4468
```

```
TukeyHSD(one_way_father_edu)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = GPA ~ father_education, data = data)
##
## $father_education
##                                                diff        lwr         upr
## High school degree-Less than high school   0.37631818 -0.1724672  0.92510355
## Some college degree-Less than high school  0.46450000 -0.1339502  1.06295018
## College degree-Less than high school       0.42920455 -0.1121140  0.97052310
## Graduate degree-Less than high school      0.14103846 -0.4156761  0.69775303
## Some college degree-High school degree     0.08818182 -0.2612378  0.43760144
## College degree-High school degree          0.05288636 -0.1858127  0.29158541
## Graduate degree-High school degree        -0.23527972 -0.5070932  0.03653376
## College degree-Some college degree        -0.03529545 -0.3728669  0.30227597
## Graduate degree-Some college degree       -0.32346154 -0.6852070  0.03828395
## Graduate degree-College degree            -0.28816608 -0.5445700 -0.03176218
##                                                p adj
## High school degree-Less than high school  0.3228592
## Some college degree-Less than high school 0.2059425
## College degree-Less than high school      0.1879975
## Graduate degree-Less than high school     0.9555990
## Some college degree-High school degree    0.9562013
## College degree-High school degree         0.9725857
## Graduate degree-High school degree        0.1229411
## College degree-Some college degree        0.9984379
## Graduate degree-Some college degree       0.1027769
## Graduate degree-College degree            0.0193294
```

Discussion: *Starting off with a discussion the plots, we see that income doesn't really tell us much. Students making between 30k and 50k have a lower median GPA than those who make between 50k and 70k, but the students that make between 15k and 30k also have higher GPAs than that group. So the results are really all over the place. We can speculate that this is because current income really isn't such a big deal for a college student. A lot of students rely on loans or sources of income such as their parents.*

Moving onto the education level of the father, we start to see some noticeable differences. Students on both extreme ends of this factor tend to have lower GPAs. Although there are only 4 observations of students whose father has less than a high school education. Of the 28 students with fathers holding graduate degrees, they have a lower GPA on average. That's interesting, and should be looked at in more detail.

Lastly, we have the student's ideal diet. These results look much like the results from student income, with no real observable differences between means for each group, because of smaller sample sizes arising from the ranges of answers they were allowed to respond in the survey. Students who chose "unclear" might have statistically lower GPAs from just being lazy, but with just 3 "unclear" responses it's unlikely that this isn't just by chance.

As promised, let's run one-way ANOVA on each factor. Income, with a p-value of .663 confirms the analysis of income above, as there's no statistically significant difference in means based on income. Father education level, with a p-value of .00715, lower than a significance of .05, has strong evidence for a difference in means between groups within this factor. Ideal diet, with a p-value of .295 comes up short of .05, showing that this factor has weak evidence for a difference in means between groups. Even with our promising "unclear" observation.

To state the above, we do need to confirm the results of our ANOVA analysis by testing for common variance (Levene's test) and normality (Shapiro-Wilkes test) for each factor. While each factor passes the common variance test, income level and ideal diet do not pass the normality test, showing strong evidence for deviating from the normal distribution. Luckily, there's a nonparametric alternative to one-way ANOVA, the Kruskal-Wallis rank sum test, which can be used when ANOVA assumptions are not met. Kruskal-Wallis confirms the null hypothesis that there is no distinct difference in means between groups for these two factors.

Finally, we can move on to our Tukey pairwise comparison for the groups within the father's education level factor. The only significant difference between groups is the difference between student's who's fathers have graduate degrees and students whose fathers have college degrees, favoring the latter. It's worth noting that this could just be by chance. You fish long enough for p-values under .05, you'll eventually find some. The best procedure would be to re-do this survey once again with a larger sample size. If I were to speculate why this is, however, maybe it's because fathers with graduate degrees might not get paid as much. Sure, getting a graduate degree might pay you more relative to your peers without them, but business majors don't get graduate degrees. If you looking to make money, chase the money. Do business. If you have real passion, get that PhD! (And maybe pay for a college statistics tutor later down the road.)