Exploring the BRFSS Data

Mohamed ESDAIRI

Setup

Load packages

```
## if you don't have forcats installed uncomment the following line:
#install.packages("forcats")

library(dplyr)
library(ggplot2)
library(knitr)
library(gridExtra)
library(forcats)
```

Load data

First we load the data set and see the dimensions:

```
load("brfss2013.RData")
dim(brfss2013)
```

```
## [1] 491775    330
```

Part 1: Data

The data for the project was collected by The Behavioral Risk Factor Surveillance System, the project collects data relative to some health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult (aged 18 years and more), non-institutionalized (people that were not at a hospital at the time of the interview) population, the data was collected by two main types of interviews: landline and cellular telephone, for the landline interview it selects a random adult from the household.

the information on the site of the project does not specify how a non response to interview is treated, also since this is a phone interview it is subject to multiple errors:

- whether or not the respondent is telling the truth, especially, and whether or not he recalls facts about his health problems and behaviors.
- the exact phrasing of the question might affect the response in a certain way.
- entry errors by the person asking the questions.

Considering all this factors, and given that the BRFSS has the resources needed to eliminate a lot of errors we can draw the following conclusion:

- The study relies on random sampling, the exact method used is Random Digit Dialing as it says on the BRFSS FAQ page: https://www.cdc.gov/brfss/about/brfss_faq.htm, so we can generalize the results to: adult (aged 18 or plus) non-institutionalized US residents.

- There is no random assignment so this is an observational study, so: we can only infer ASSOCIATION and NOT CAUSATION.

Part 2: Research questions

As a college student, in my analysis I will try to explore the associations that exist between some of the thing I care about like some life habits (exercise, diet and sleep) and how they relate to some health problems (overweight,high blood pressure and heart attack) also one of the most important issues is how can education level be associated with income level.

Research quesion 1: Some studies suggest that exercise and healthy food consumption(especially fruit and vegetables) help to prevent overweight, how can we describe the associations between physical activity and overweight, and between healthy food consumption and overweight?

Research quesion 2: Sleep is essential to humans, and most studies say that 7 to 9 hours is the correct amount of sleep that every one should get. It is common to link sleep deprivation to some health problems such an increasing risk of heart attack and high blood pressure, so is there any association between sleep deprivation (less than 7h) and high blood pressure and also heart attack?

Research quesion 3: How does education level associate with income level?

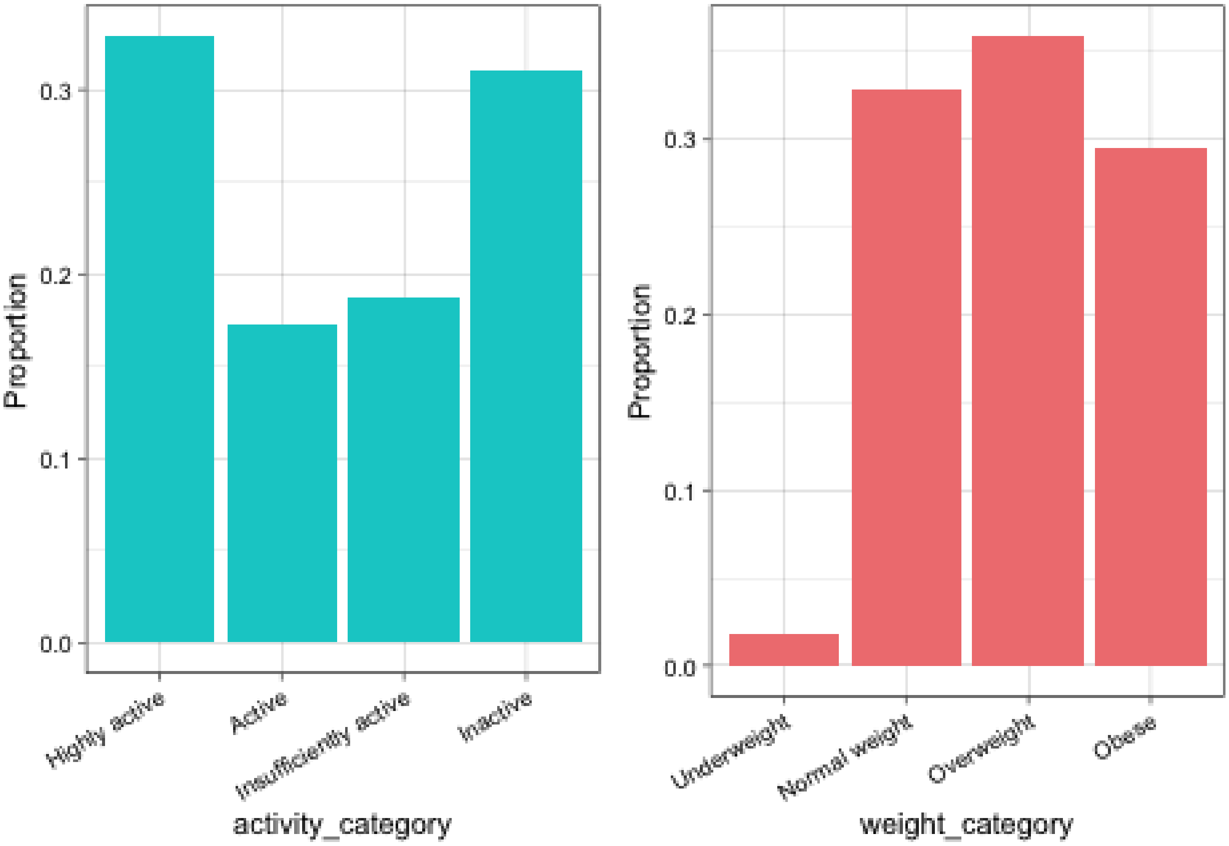
Part 3: Exploratory data analysis

In this part we will use a bunch of variables to verify if the associations suggested in the research questions hold.

Research quesion 1:

Let us start by exploring two variables we will use in the analysis in the first question, “Physical Activity Categories” (X_pacat1) and Computed Body Mass Index Categories (X_bmi5cat) both of them are categorical variables.

```
brfss2013 %>%
  select(X_pacat1, X_bmi5cat) %>%
  na.omit() ->
  weight_by_activity

  names(weight_by_activity) <- c("activity_category", "weight_category")

  create_partition_barplot <- function(data, x, fill_color, label_angle) {
    ggplot(data = data, aes(x = get(x))) +
        geom_bar(aes(y =  (..count..)/sum(..count..)), fill = fill_color) +
        theme_linedraw() +
        theme(axis.text.x = element_text(angle = label_angle, hjust = 1)) +
        labs(x = x, y = "Proportion")
  }


  activity_plot <-
    create_partition_barplot(weight_by_activity, "activity_category", "cyan3", 30)


  body_mass_index_plot <-
    create_partition_barplot(weight_by_activity, "weight_category", "lightcoral", 30)
  grid.arrange(activity_plot, body_mass_index_plot, ncol=2)
```



We can see that the Physical Activity Categories can take one of 4 values: Highly active, Active, Insufficiently active and Inactive. The most common category among observations is Highly active, followed by Inactive then Insufficiently active and for the last category Active. Computed Body Mass Index Categories takes 4 values Underweight, Normal weight, Overweight and Obese. most observations fall under Overweight followed by Normal weight then Obese and for the last category Underweight.

Now we will prepare the data to plot the variables in respect to each other, let us start by preparing the data: lets group the data by the activity_category and then by weight_category and take the counts under each category and sub-category, we will achieve that using the following chunk of code:

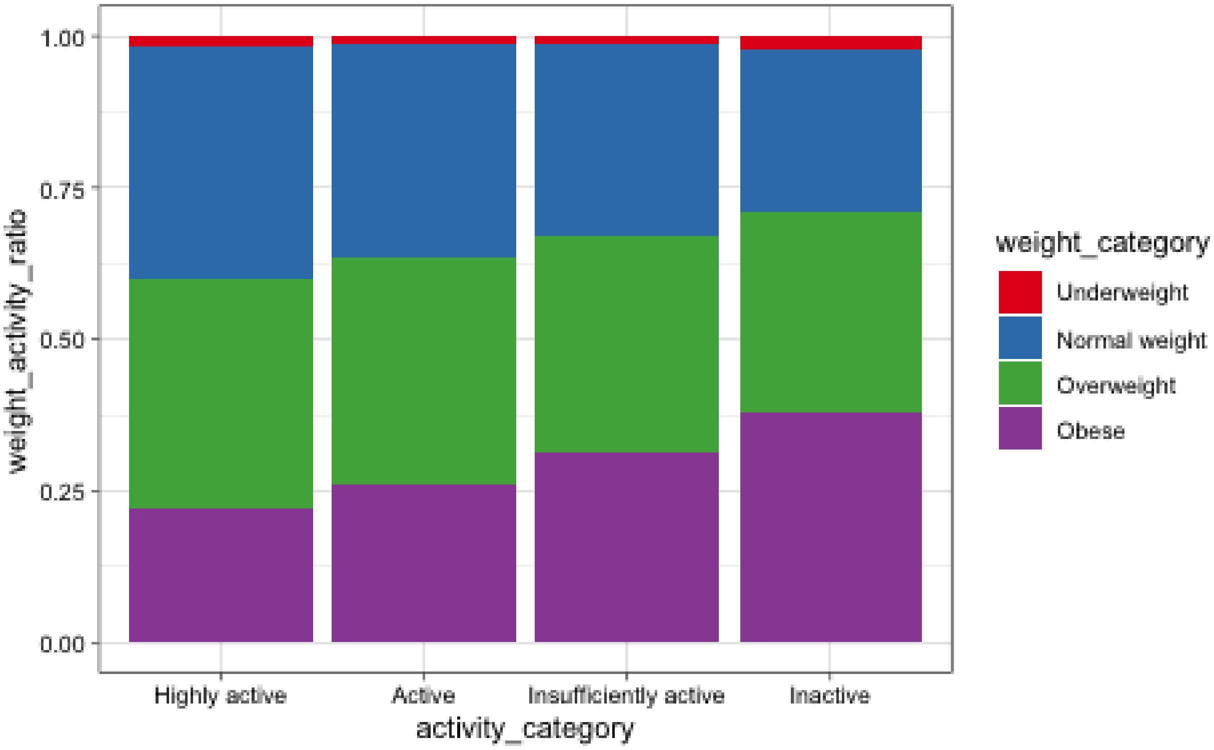
```
  weight_by_activity %>%
  group_by(activity_category, weight_category ) %>%
  summarise(activity_counts = length(activity_category),
            weight_counts = length(weight_category)) %>%
  group_by(activity_category) %>%
  mutate(activity_counts = sum(weight_counts)) %>%
  ungroup() ->
  weight_by_activity_type_data
```

we will take the proportion of each sub-category in respect to each parent category using this command:

```
weight_by_activity_type_data$weight_activity_ratio <-
  round(weight_by_activity_type_data$weight_counts /
        weight_by_activity_type_data$activity_counts, 3)
```

lets plot the results:

```
plot_weight_activity <-
  qplot(x = activity_category,
      y = weight_activity_ratio,
      fill = weight_category,
      data = weight_by_activity_type_data,
      geom = 'col') +
  scale_fill_brewer(palette = 'Set1') +
  theme_linedraw()

  print(plot_weight_activity)
```



As we can see in this plot the physical activity is associated with the weight category, since the obesity rates increase when the physical activity decreases, and the normal weight proportion drops with the drop of physical activity, this proves association between the two variables Physical Activity Categories and Computed Body Mass Index Categories but we can't conclude causation since this is an observational study.

Now we will see how both fruit and vegetable consumption affect the BMI(body mass index), we will add some new variables to analysis, we are talking here about: X_bmi5: Computed Body Mass Index this is a numerical continuous variable X_vegesum: Total Vegetables Consumed Per Day this is a numerical continuous variable X_frutsum: Total Fruits Consumed Per Day this is a numerical continuous variable

we will also keep the other variables: X_pacat1, X_bmi5cat so we can later combine the result and see how the combination of variables related to healthy food and physical activity affect body weight, we will start by extracting the relevant variables and applying some transformation and calculating some new variables:

```
  brfss2013 %>%
  select(X_frutsum, X_vegesum, X_pacat1, X_bmi5, X_bmi5cat) %>%
  na.omit() ->
  weight_by_healthy_food_and_activity

  weight_by_healthy_food_and_activity$frut_veg_sum <-
```

```
    weight_by_healthy_food_and_activity$X_frutsum +
    weight_by_healthy_food_and_activity$X_vegesum




  weight_by_healthy_food_and_activity %>%
  select(frut_veg_sum, X_pacat1, X_bmi5, X_bmi5cat) ->
  weight_by_healthy_food_and_activity


  weight_by_healthy_food_and_activity$X_bmi5 <-
    weight_by_healthy_food_and_activity$X_bmi5/100

  weight_by_healthy_food_and_activity$frut_veg_sum <-
    weight_by_healthy_food_and_activity$frut_veg_sum/100

  names(weight_by_healthy_food_and_activity) <-
    c("total_frut_veg_cons", "actv_category", "BMI", "weight_category")
```

a quick look at the summary statistics for both variables:

```
weight_by_healthy_food_and_activity %>%
  select(total_frut_veg_cons, BMI) %>%
  summarize_all(funs(mean, sd)) ->
  weight_by_healthy_food_and_activity_summary


weight_by_healthy_food_and_activity %>%
  select(total_frut_veg_cons, BMI) %>%
  summarize_all(funs(median, IQR)) ->
  weight_by_healthy_food_and_activity_robust_summary



  kable(weight_by_healthy_food_and_activity_summary)
```

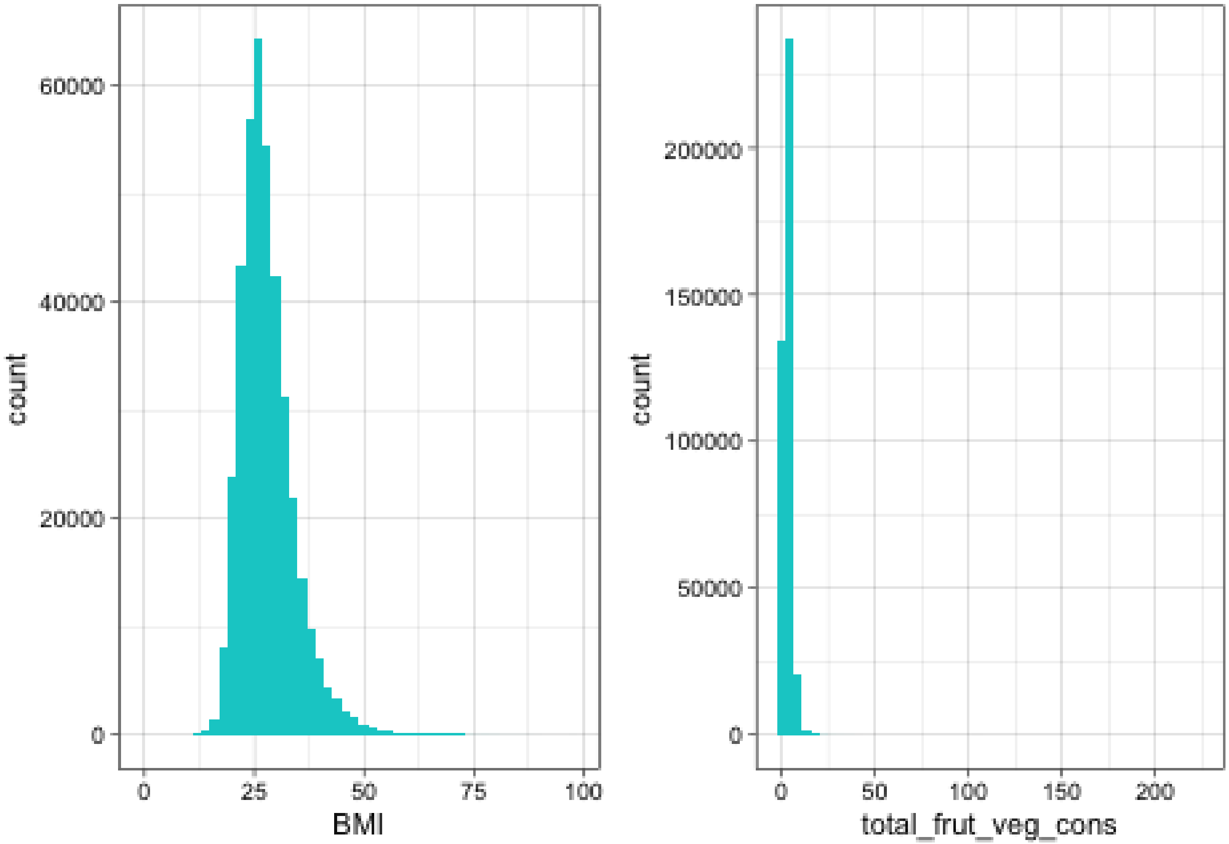
total_frut_veg_cons_mean	BMI_mean	total_frut_veg_cons_sd	BMI_sd
3.288818	27.90645	2.433948	6.197378

```
  kable(weight_by_healthy_food_and_activity_robust_summary)
```

total_frut_veg_cons_median	BMI_median	total_frut_veg_cons_IQR	BMI_IQR
2.89	26.83	2.3	7.19

Lets take a look at the distribution of the variables:

```
evalutate_bmi_and_frut_veg_distributions <- function(weight_by_healthy_food_and_activity_data) {

  BMI_hsitogram <-
    ggplot(data = weight_by_healthy_food_and_activity_data) +
      geom_histogram(aes(x = BMI),
                     bins = 50,
                     fill='cyan3') +
    theme_linedraw()
  frut_veg_consum_histogram <-
    ggplot(data = weight_by_healthy_food_and_activity_data) +
      geom_histogram(aes(x = total_frut_veg_cons),
                     bins = 50,
                     fill='cyan3') +
    theme_linedraw()

  data_dist_histogam_plots <-
    grid.arrange(BMI_hsitogram, frut_veg_consum_histogram, ncol=2)

 data_dist_histogam_plots
}

evalutate_bmi_and_frut_veg_distributions(weight_by_healthy_food_and_activity)
```

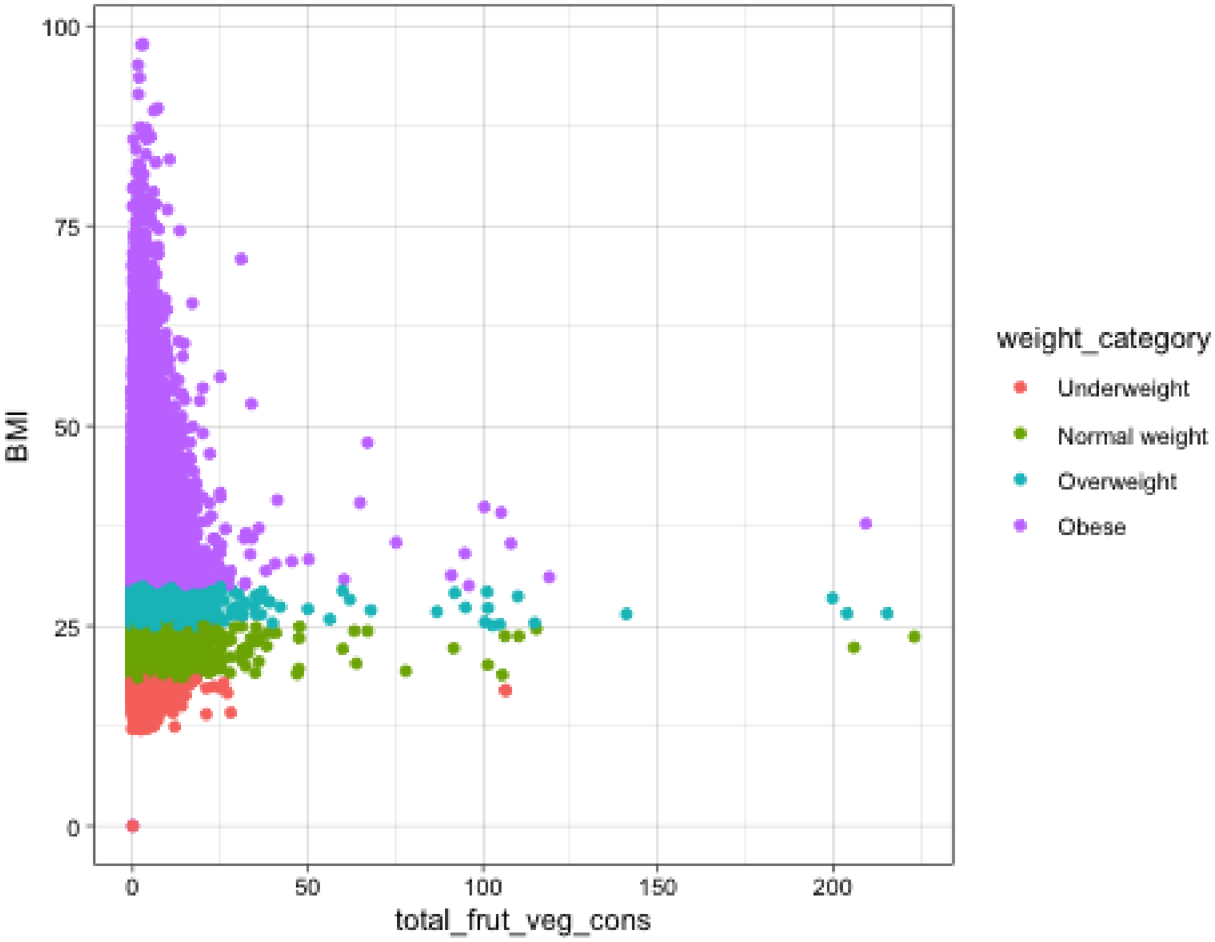


```
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

we can see that both distributions are right skewed, but that the distribution of total fruit and vegetable consumption is more skewed than the BMI distribution.

We will plot the BMI against the fruit and vegetable consumption, to see if there is any prominent relationship between the two variables.

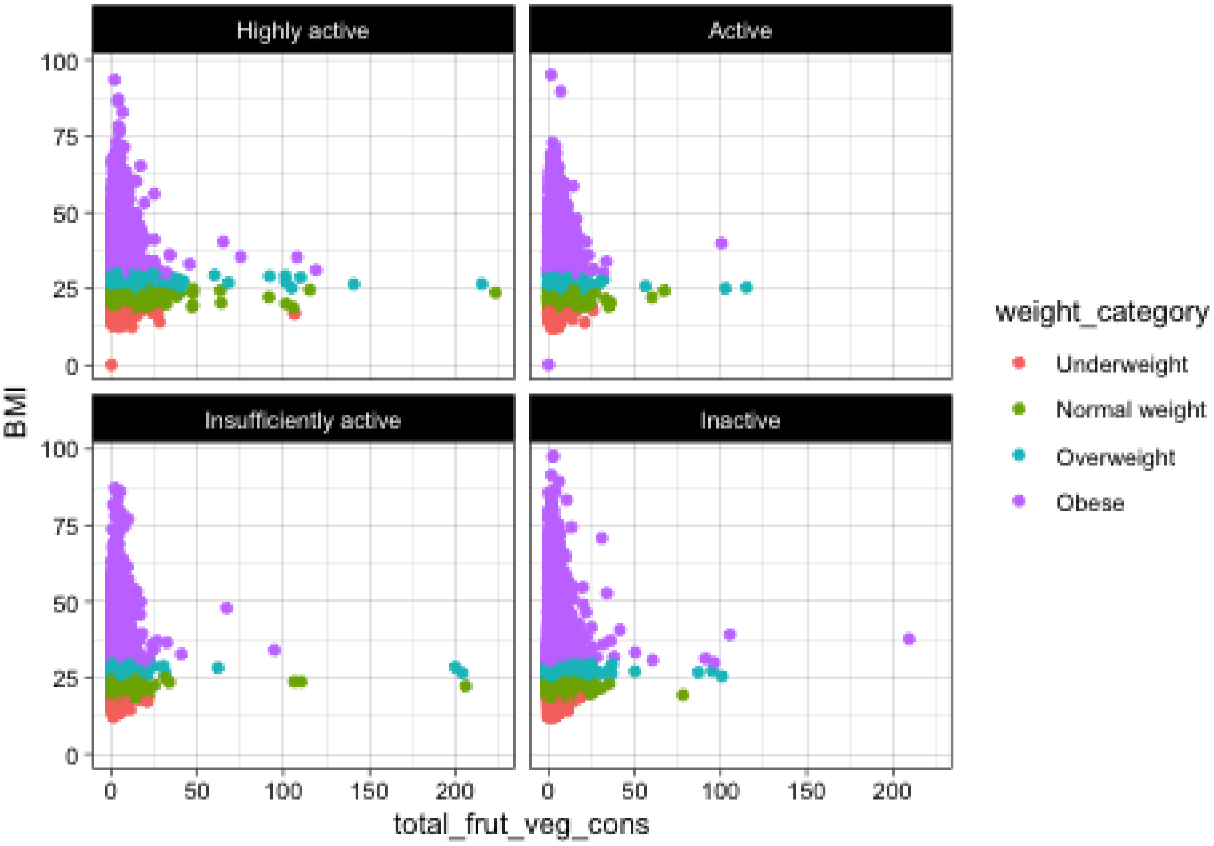
```
ggplot(data=weight_by_healthy_food_and_activity,
       aes(x=total_frut_veg_cons, y=BMI)) +
  geom_point(aes(color=weight_category)) +
  theme_linedraw()
```



we can see that as the consumption of vegetables and fruits increases the number of observations that exhibit obesity and underweight decreases and the BMI tends to go in the direction of a healthy BMI, but other than this somewhat unstable association we can't conclude anything else about the association, so we will try to understand how both variables: fruit/vegetable consumption and physical activity affect the BMI and BMI category.

so let us create a plot that the 4 variables:

```
ggplot(data=weight_by_healthy_food_and_activity,
       aes(x=total_frut_veg_cons, y=BMI)) +
  geom_point(aes(color=weight_category)) +
  scale_shape_identity() +
  theme_linedraw() +
  facet_wrap(~actv_category)
```



It is clear that for all people who are at least Insufficiently active, as the fruit and vegetable consumption increase the BMI tends to go towards the interval [18.5 ; 25] which is the normal weight interval, and also that the observations that show obesity and overweight tends to shrink with the increase of vegetable and fruit consumption. In the other hand for people who are Inactive even though the increase of healthy food (veg and fruit) tends to make the BMI go towards the normal weight interval some observations still exhibit obesity and overweight for high levels of healthy food consumption.

conclusion: To answer the First question we can safely say that there is a strong association between physical activity and body weight on one hand, also there is a somewhat weak association between healthy food consumption and body weight. But the strongest association is between (physical activity + healthy food consumption) and body weight. We can't infer any causal connection since this is not an experiment and there is no control and treatment groups.

Research quesion 2:

Now we will if sleep is associated to heart attack and high blood pressure, we will use the following variables: sleptim1: How Much Time Do You Sleep, this is a numerical continuous variable that indicates the average amount of time a person get in a 24h time period. cvdinfr4: Ever Diagnosed With Heart Attack, this is a categorical variable that indicates whether or not a person has ever been diagnosed with heart attack. X_rfhype5: High Blood Pressure Calculated Variable, this is a categorical variable that indicates if the person has ever been told he has high blood pressure by any health professional.

we will calculate a new variable sleep_condition, that will contain sleep deprived if sleptim1 is less than 7 hours and normal otherwise normal.

```
brfss2013 %>%
  select(sleptim1, cvdinfr4, X_rfhype5) %>%
  na.omit() %>%
  mutate(sleep_condition = ifelse(sleptim1 < 7, "sleep deprived", "normal")) %>%
  select(sleep_condition, cvdinfr4, X_rfhype5)->
  heart_attack_blood_pressure_by_sleep

  names(heart_attack_blood_pressure_by_sleep) <-
    c("sleep_condition", "heart_attack", "high_blood_pressure")

  heart_attack_blood_pressure_by_sleep$high_blood_pressure <-
    fct_rev(heart_attack_blood_pressure_by_sleep$high_blood_pressure)
```

we will explore the partitions of the categories of all the 3 variables:

```
heart_attack_blood_pressure_by_sleep %>%
  group_by(sleep_condition) %>%
  summarise(m = n()) %>%
  mutate(sleep_condition_pecentage = round(100 * m/sum(m), 2)) %>%
  select(sleep_condition, sleep_condition_pecentage) ->
  sleep_condition_partitions

heart_attack_blood_pressure_by_sleep %>%
  group_by(heart_attack) %>%
  summarise(m = n()) %>%
  mutate(heart_attack_percentage = round(100 * m/sum(m), 2)) %>%
  select(heart_attack, heart_attack_percentage) ->
  heart_attack_partitions

heart_attack_blood_pressure_by_sleep %>%
  group_by(high_blood_pressure) %>%
  summarise(m = n()) %>%
  mutate(high_blood_pressure_percentage = round(100 * m/sum(m), 2)) %>%
  select(high_blood_pressure, high_blood_pressure_percentage) ->
  high_blood_pressure_partitions

kable(sleep_condition_partitions)
```

sleep_condition	sleep_condition_pecentage
normal	67.27
sleep deprived	32.73

```
kable(heart_attack_partitions)
```

heart_attack	heart_attack_percentage
Yes	5.91
No	94.09

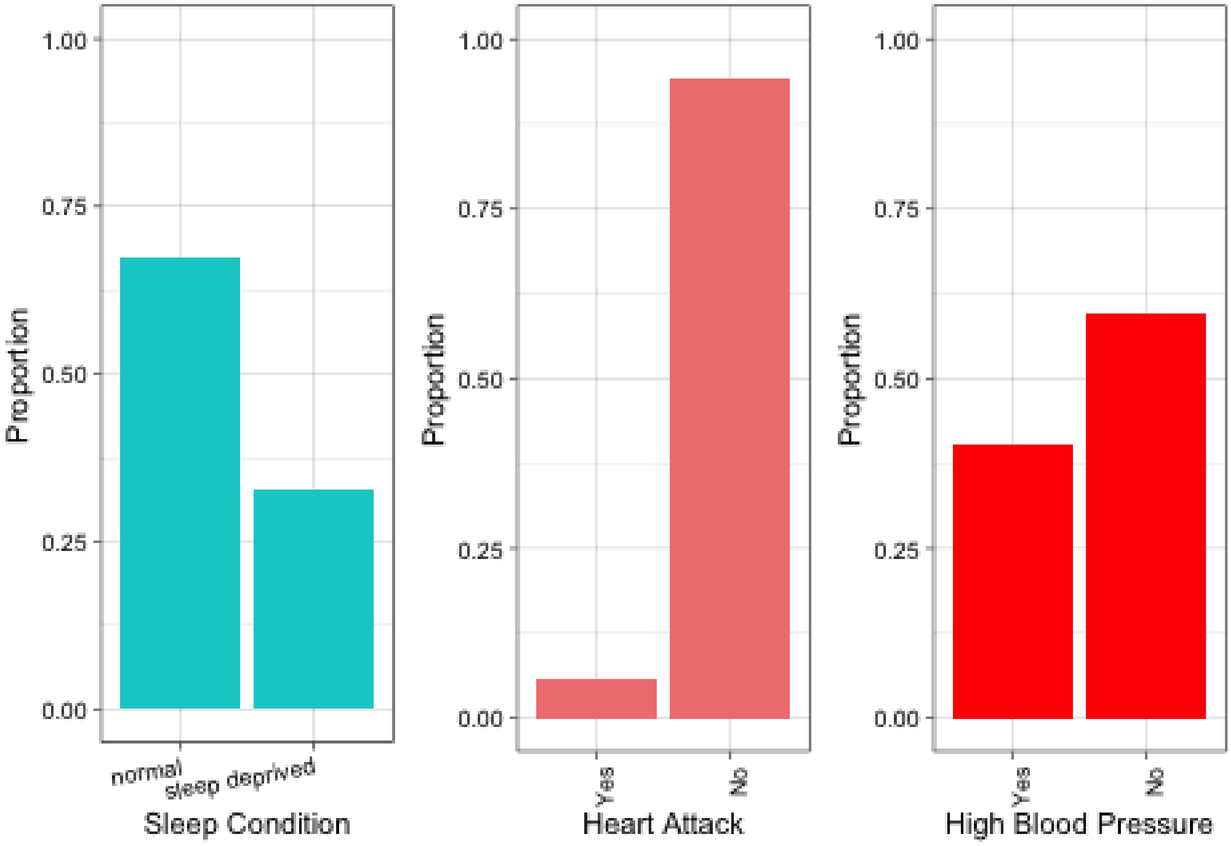
```
kable(high_blood_pressure_partitions)
```

high_blood_pressure	high_blood_pressure_percentage
Yes	40.22
No	59.78

We can clearly see that 67.27% of the population get at least 7 hours of sleep in a 24h period. also we can see that heart attack is a relatively rare disease since only about 5.91% of the population exhibit this condition, if we compare it to high Blood pressure we can see that for a random person, it is almost 4 times more likely to have high blood pressure that it is to have heart attack.

we will plot those percentages on the same the scale to take a look and compare them:

```
sleep_condition_plot <-
  ggplot(heart_attack_blood_pressure_by_sleep, aes(x = sleep_condition)) +
        geom_bar(aes(y =  (..count..)/sum(..count..)), fill='cyan3') +
        theme_linedraw() +
        theme(axis.text.x = element_text(angle = 10, hjust = 1)) +
        labs(x = "Sleep Condition", y = "Proportion") +
        coord_cartesian(ylim = c(0, 1))

  heart_attack_plot <-
  ggplot(heart_attack_blood_pressure_by_sleep, aes(x = heart_attack)) +
        geom_bar(aes(y =  (..count..)/sum(..count..)),  fill='lightcoral') +
        theme_linedraw() +
        theme(axis.text.x = element_text(angle = 90, hjust = 1 )) +
        labs(x = "Heart Attack", y = "Proportion") +
        coord_cartesian(ylim = c(0, 1))

  high_blood_pressure_plot <-
  ggplot(heart_attack_blood_pressure_by_sleep, aes(x = high_blood_pressure)) +
        geom_bar(aes(y =  (..count..)/sum(..count..)),  fill='red') +
        theme_linedraw() +
        theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
        labs(x = "High Blood Pressure", y = "Proportion") +
        coord_cartesian(ylim = c(0, 1))



  grid.arrange(sleep_condition_plot, heart_attack_plot, high_blood_pressure_plot, ncol=3)
```



we can see that on the same scale there is a big difference in the partition of high blood pressure and heart attack.

to explore the relation between those variables we will make contingency tables of heart attack in respect to sleep condition and then of high blood pressure in respect to sleep condition:

```
rounded_prop_tables <- function(first_factor, second_factor)
{

    round(prop.table(table(first_factor, second_factor))*100, 2)

}
```

Heart Attack:

```
heart_attack_cont_table <-
  rounded_prop_tables(heart_attack_blood_pressure_by_sleep$sleep_condition,
                      heart_attack_blood_pressure_by_sleep$heart_attack)

kable(heart_attack_cont_table)
```

	Yes	No
normal	3.72	63.56
sleep deprived	2.19	30.53

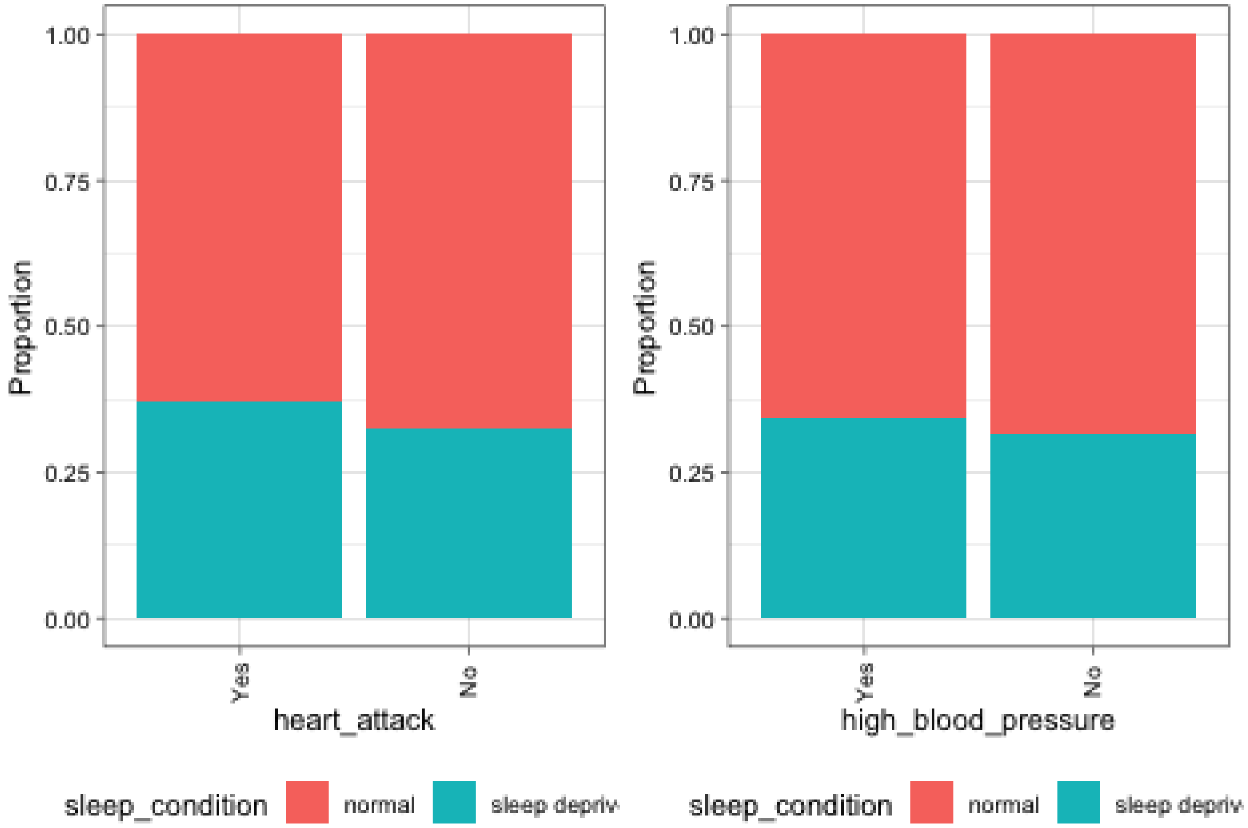
High Blood Pressure:

```
high_blood_pressure_cont_table <-
  rounded_prop_tables(heart_attack_blood_pressure_by_sleep$sleep_condition,
                      heart_attack_blood_pressure_by_sleep$high_blood_pressure)

kable(high_blood_pressure_cont_table)
```

	Yes	No
normal	26.45	40.82
sleep deprived	13.77	18.95

we can't tell from those tables if there is any association between heart attack and high blood pressure, trying a visual representation may give us some clues about this association:

```
create_stacked_barplot <- function(data, x, fill ) {

    ggplot(data = data) +
    aes(x = get(x), fill = get(fill)) +
    geom_bar(position = "fill") +
    theme_linedraw() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5),
                 legend.position = "bottom") +
    labs(x = x, y = "Proportion") +
    guides(fill=guide_legend(title=fill))

}


plot_heart_attack_by_sleep <-
  create_stacked_barplot(heart_attack_blood_pressure_by_sleep,
                         "heart_attack",
                         "sleep_condition")


plot_high_blood_pressure_by_sleep <-
  create_stacked_barplot(heart_attack_blood_pressure_by_sleep,
                         "high_blood_pressure",
                         "sleep_condition")



grid.arrange(plot_heart_attack_by_sleep,
             plot_high_blood_pressure_by_sleep,
             ncol=2)
```



we can see that heart attack it is more likely for people who are sleep deprived and this is also the case for high blood pressure but we can see that this is a slight difference, it might as well be due to chance, so to test this we will do some simulations.

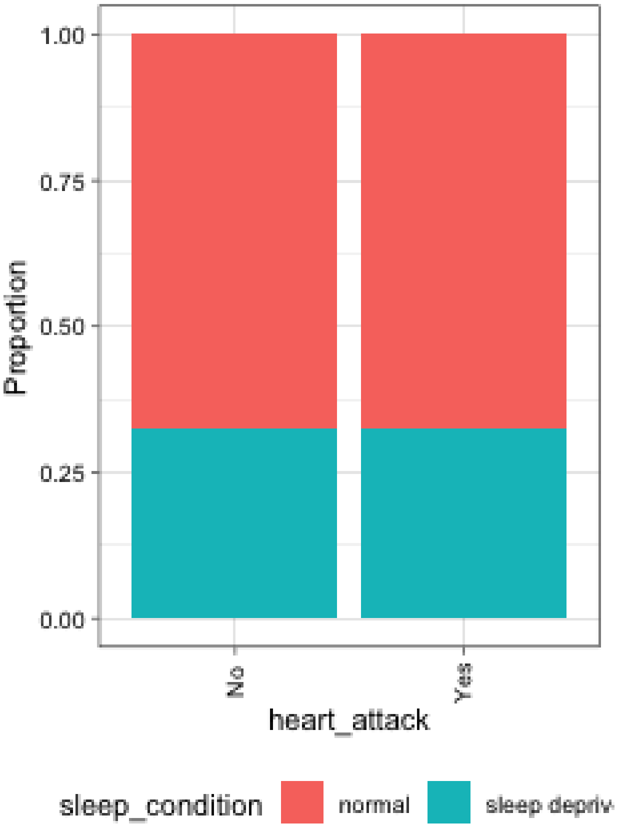
This code will generate a random data set based on the percentages of each category, and we will examine how this compares to our data, to this generated from the underline distribution and see if the difference that we noticed in the previous section is due to chance or if it is related to the amount of sleep that a person gets.

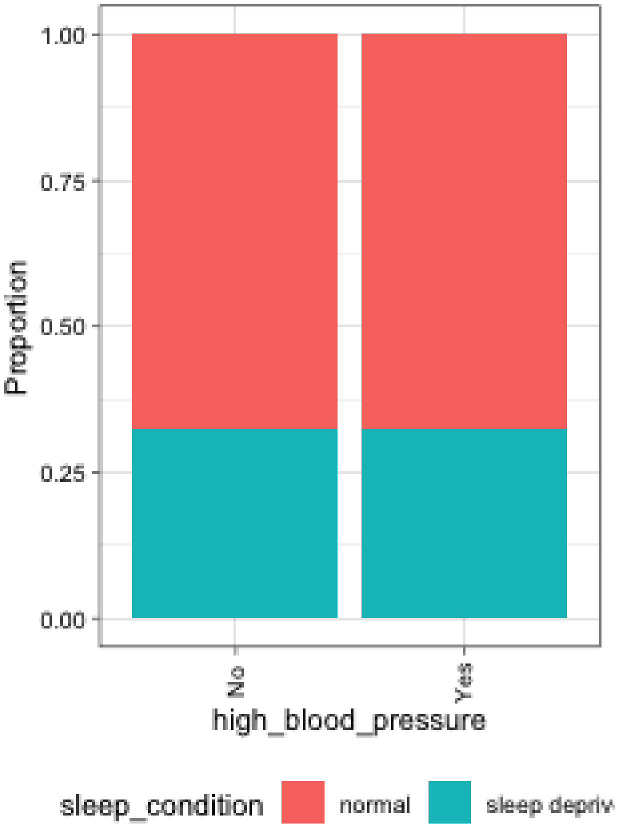
```
create_factor_simulation <- function(size, first_factor,
                                     first_prob,
                                     second_factor,
                                     second_prob,
                                     replace) {


  as.factor(sample( c(first_factor, second_factor),
                    size, prob = c(first_prob,  second_prob),
                    replace = replace))

}


concatenate_simulations <- function(simulation_arguments) {

replace_simulation = TRUE
replace_concatenation  = FALSE

sleep_condition_simulation <-
```

```
 create_factor_simulation(simulation_arguments[["sleep_condition", "sim_size"]],
                          simulation_arguments[["sleep_condition", "first_factor"]],
                          simulation_arguments[["sleep_condition", "first_prob"]],
                          simulation_arguments[["sleep_condition", "second_factor"]],
                          simulation_arguments[["sleep_condition", "second_prob"]],
                          replace_simulation)

heart_attack_simulation <-
create_factor_simulation(simulation_arguments[["heart_attack", "sim_size"]],
                          simulation_arguments[["heart_attack", "first_factor"]],
                          simulation_arguments[["heart_attack", "first_prob"]],
                          simulation_arguments[["heart_attack", "second_factor"]],
                          simulation_arguments[["heart_attack", "second_prob"]],
                          replace_simulation)

high_blood_pressure_simulation <-
 create_factor_simulation(simulation_arguments[["high_blood_pressure", "sim_size"]],
                          simulation_arguments[["high_blood_pressure", "first_factor"]],
                          simulation_arguments[["high_blood_pressure", "first_prob"]],
                          simulation_arguments[["high_blood_pressure", "second_factor"]],
                          simulation_arguments[["high_blood_pressure", "second_prob"]],
                          replace_simulation)
data.frame(

 sleep_condition = sleep_condition_simulation[sample(length(sleep_condition_simulation),
                                              length(sleep_condition_simulation),
                                              replace = replace_concatenation)
                                       ],



heart_attack = heart_attack_simulation[sample(length(heart_attack_simulation),
                                              length(heart_attack_simulation),
                                              replace = replace_concatenation)
                                       ],

high_blood_pressure = high_blood_pressure_simulation[sample(length(high_blood_pressure_simulation),
                                              length(high_blood_pressure_simulation),
                                              replace = replace_concatenation)
                                       ]

)

}



 simulation_arguments <- t( data.frame(

     c(480818, "normal", 0.6727, "sleep deprived", 0.3273),
      c(480818, "Yes", 0.0591, "No", 0.9409),
      c(480818, "Yes",  0.4022,  "No", 0.5978)
    ))
 colnames(simulation_arguments) <-
   c("sim_size",
     "first_factor",
     "first_prob",
     "second_factor",
     "second_prob")

 row.names(simulation_arguments) <-
   c("sleep_condition",
     "heart_attack",
     "high_blood_pressure")
```

Lets generate some data and plot it to make it easier to see if there is any difference:

```
simulated_heart_attack_blood_pressure_by_sleep <-
  concatenate_simulations(simulation_arguments)

plot_heart_attack_by_sleep <-
  create_stacked_barplot(simulated_heart_attack_blood_pressure_by_sleep,
                         "heart_attack",
                         "sleep_condition")


plot_high_blood_pressure_by_sleep <-
  create_stacked_barplot(simulated_heart_attack_blood_pressure_by_sleep,
                         "high_blood_pressure",
                         "sleep_condition")


sleep_plots <-
  grid.arrange(plot_heart_attack_by_sleep,
               plot_high_blood_pressure_by_sleep,
               ncol=2)
```





we can generate random data as many times as we want, but this does not show any difference between heart attack and blood pressure percentages for people who have enough sleep and those who are sleep deprived, so the difference that exhibits the original data can NOT be due to chance.

conclusion: From all those analyses we can conclude both heart attack and high blood pressure have an association with weather or not a person gets enough sleep, this just an ASSOCIATION, to infer causation we need to conduct further experiments with control and treatment groups and random assignment.

Research quesion 3:

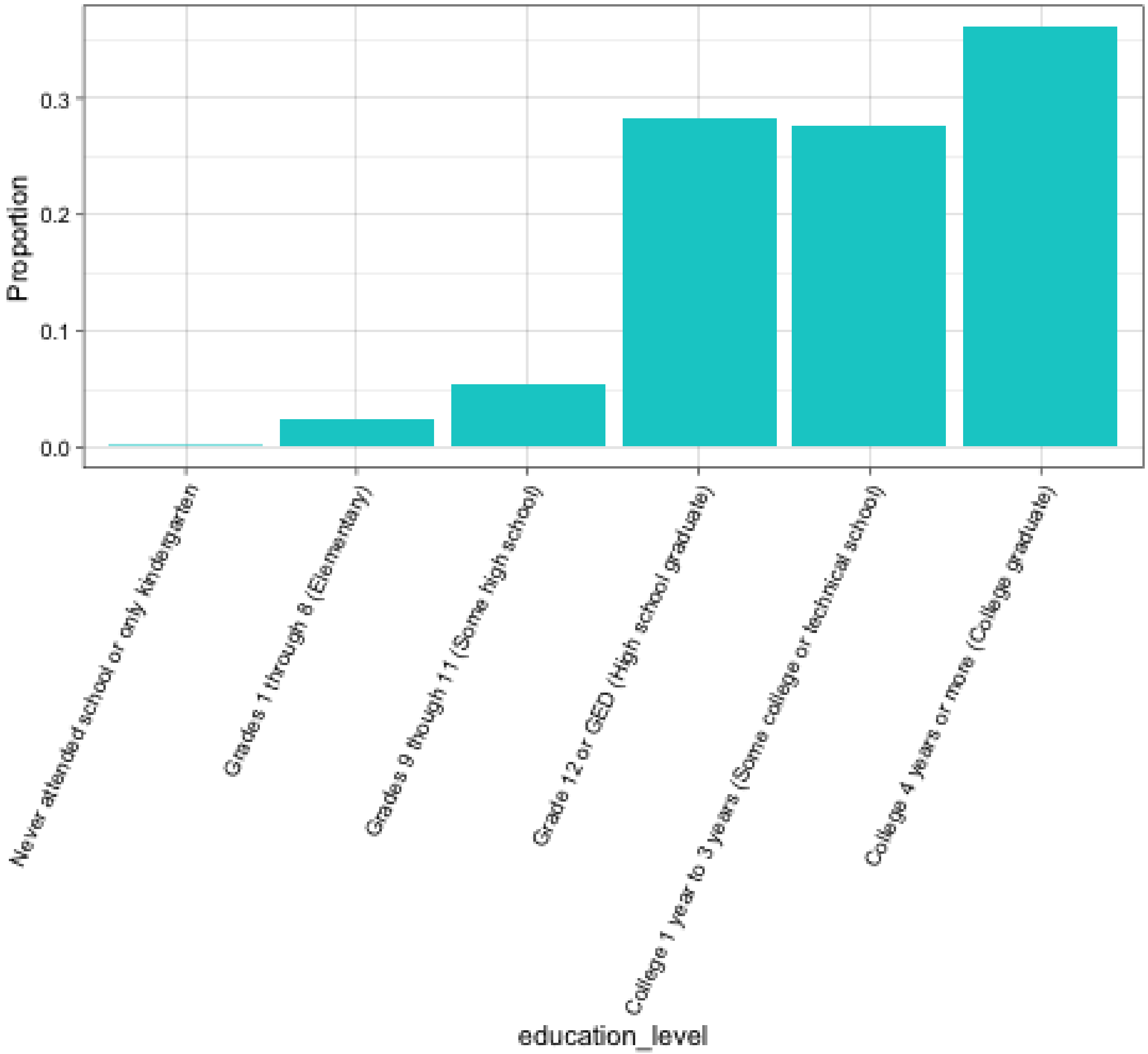
Now we will see if there is any association between education level and income level, we will use a set of two categorical variables: educa: Education Level income2: Income Level

Let's first extract the data that we will work with and remove all the missing values:

```
brfss2013 %>%
  select(educa, income2) %>%
  na.omit() ->
  income_by_education

  names(income_by_education) <- c("education_level", "income_category")
```

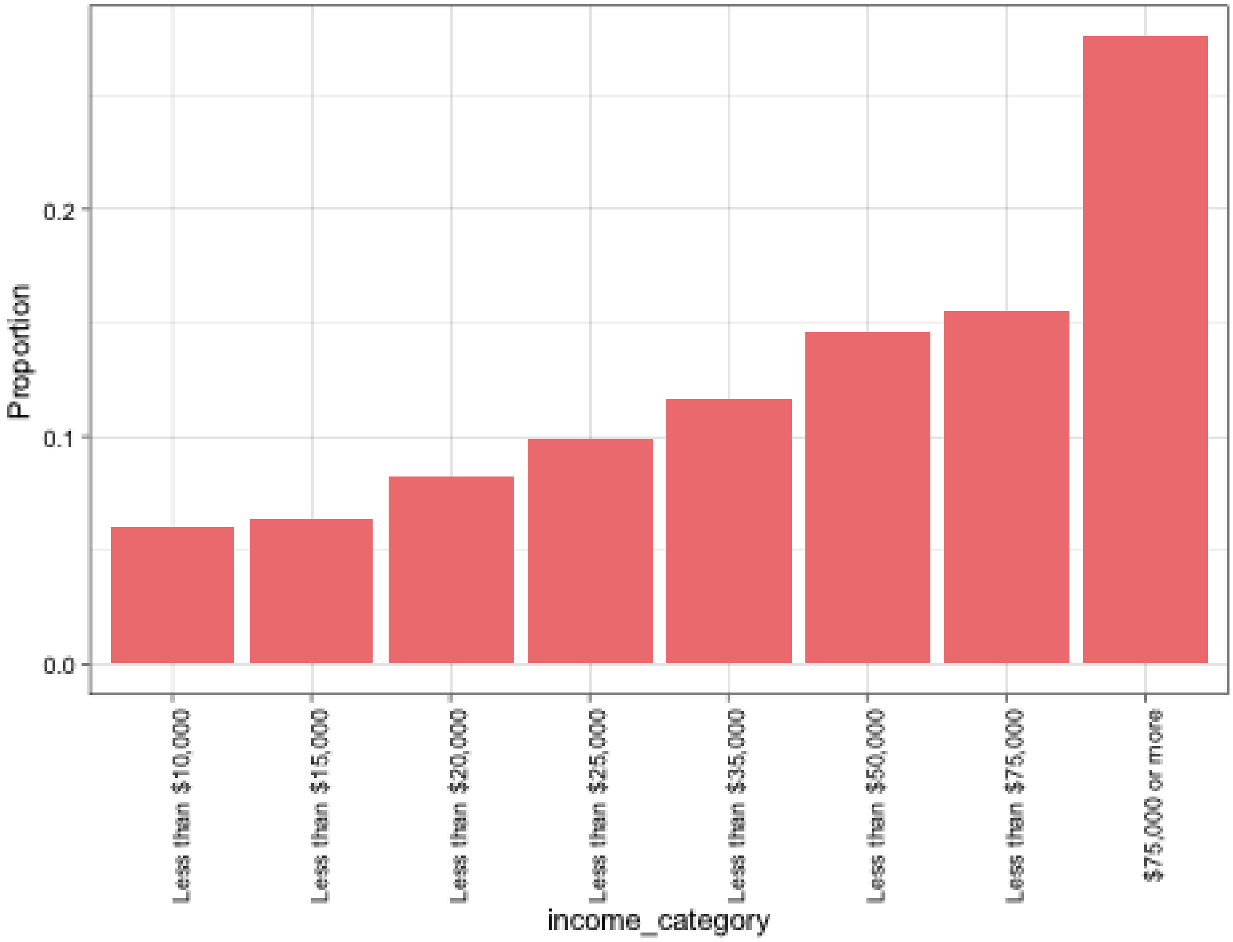
we will take a look at the proportions of the variables and how the observations span over each of the categories:

```
create_partition_barplot(income_by_education, "education_level", "cyan3", 65)
```



we can see that there are 6 categories in education level variable, and we can see clearly that the majority (more than 80%) of people in the data set have at least graduated high school.

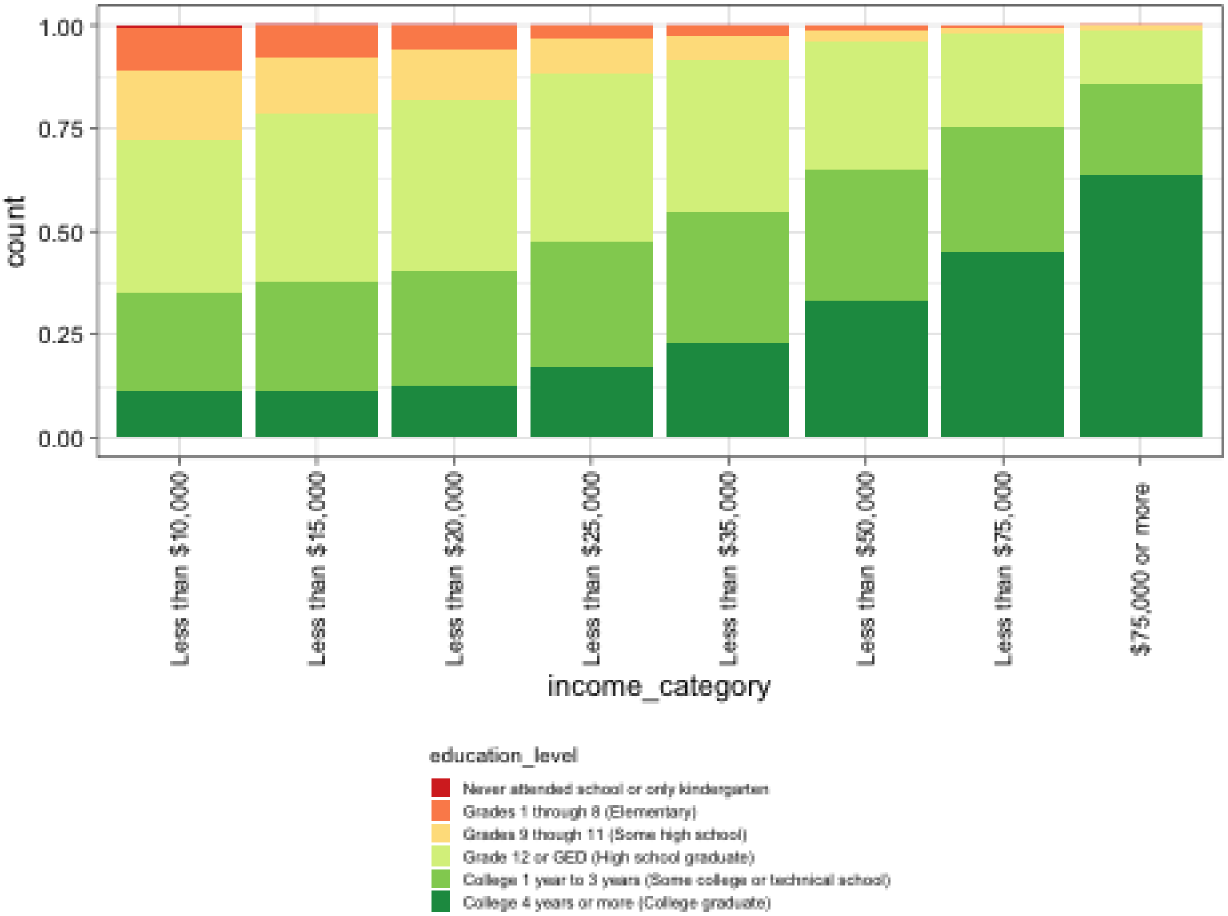
```
create_partition_barplot(income_by_education, "income_category", "lightcoral", 90)
```



The income level is comprised of 8 categories, and we can notice that most observations in the population: more than 60% have an income that exceeds 25k$.

Now that we have an idea about the structure of the variables that we'll be using in our analysis, let's plot them against each other and see what relationship will present itself:

```r
plot_income_by_education <-
  ggplot(data = income_by_education) +
  aes(x = income_category, fill = education_level) +
  geom_bar(position = "fill") +
  theme_linedraw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5),
             legend.position = "bottom",
             legend.direction = "vertical",
             plot.title = element_text(size = 12, face = "bold"),
             legend.title=element_text(size=8),
             legend.text=element_text(size=6),
             legend.key.size = unit(0.3, "cm")) +
  coord_fixed(ratio = 3) +
  scale_fill_brewer(palette = 'RdYlGn')


print(plot_income_by_education)
```



in the previous plot we can see that as we move up in income levels the percentages of people with higher levels of education tends to grow, and those with education level less than high school diploma tends to shrink, the percentages of people who have attended college for less than 4 years is pretty much constant across all income levels except for income above 75k$ where it shrinks a little.

conclusion: so we can observe a positive association between Income level and Education level in the data and we can generalize this to the population of interest because this is a random sample, causation can NOT be inferred since this is an observational study.