Exploring the BRFSS data

## Setup

### Load packages

library(ggplot2)  
library(dplyr)

### Load data

load("brfss2013")

## Part 1: Data

**Type of study**: Prospective Observational (Survey) because the data was collected throughout the study. Since the data was collected by random sampling, the dataset can be used to make decisions that generalise well. However, no causal relationship can be established between variables.

**Sampling method:** Disproportionate Stratified Sampling has been used. Telephone numbers have been startified into two lists (high density or medium density) based on the presumed density of known telephone household numbers.

**Possible Implications:** Given that the interviews are time-consuming and that there are many optional modules in the questionnaire, the sampling may have suffered from voluntary bias.

**Dimensions of the dataset:** 491775 rows \* 330 columns

## Part 2: Research questions

**Research quesion 1:** Studying the relation between the body mass index and educational/ income categories.

* The intention is to explore whether obesity/ overweight is prominent among respondents with high incomes. Many a time affluence is observed to negatively impact health because of a mindset of leading “lavish” lifestyles. Is it really true?
* The correlation between education and BMI will help to establish whether there is a general awareness among those with higher educational qualifications, because in addition to subject specific knowledge, the tenets of leading a healthy life should not be foregone.

**Research quesion 2:** Studying the impact of green vegetables on cardiovascular ailments.

**Research quesion 3:** Factors affecting physical activity of respondents.

* These may include smoking and drinking habits. The purpose is to investigate whether smoking and drinking are recreational or are affecting the tendency of respondents to do longer hours of physical activity

## Part 3: Exploratory data analysis

**Research quesion 1:**

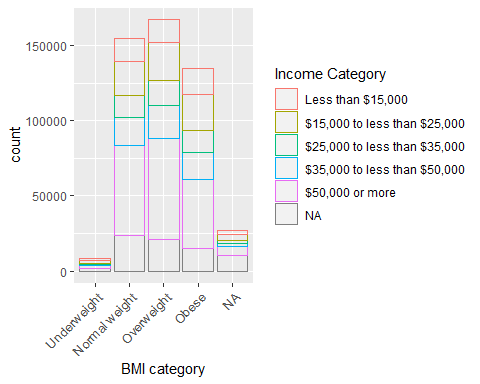
brfss2013 %>%   
 group\_by(X\_incomg) %>%  
 summarise(median\_bmi = median(X\_bmi5, na.rm = TRUE))

## Warning: Factor `X\_incomg` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

## # A tibble: 6 x 2  
## X\_incomg median\_bmi  
## <fct> <dbl>  
## 1 Less than $15,000 2744  
## 2 $15,000 to less than $25,000 2732  
## 3 $25,000 to less than $35,000 2712  
## 4 $35,000 to less than $50,000 2720  
## 5 $50,000 or more 2658  
## 6 <NA> 2609

**Observation:** The median BMI is highest for those who have an income of less than $15000. It is the lowest for those who have an income of more than $50000. Otherwise, the median is roughly the same across each income group.

ggplot(data = brfss2013)+  
 labs(x = "BMI category", color = "Income Category")+  
 geom\_bar(mapping = aes(x = X\_bmi5cat, color = X\_incomg), fill = NA)+theme(axis.text.x = element\_text(size = 10, angle = 45, hjust = 1))



**Observation:** The segmented barplot, gives us a more informative analysis. It is to be noted that majority of obese and overweight people have an income of more than $50000.

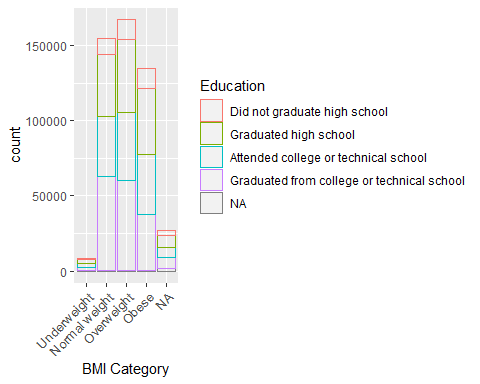
brfss2013 %>%   
 group\_by(X\_educag) %>%  
 summarise(median\_bmi = median(X\_bmi5, na.rm = TRUE))

## Warning: Factor `X\_educag` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

## # A tibble: 5 x 2  
## X\_educag median\_bmi  
## <fct> <dbl>  
## 1 Did not graduate high school 2750  
## 2 Graduated high school 2734  
## 3 Attended college or technical school 2720  
## 4 Graduated from college or technical school 2589  
## 5 <NA> 2582

**Observation:** There is a clear correlation to be observed. Respondents with higher educational qualifications tend to have lower BMIs. This may be a result of better awareness.

ggplot(data = brfss2013)+  
 labs(x = "BMI Category", color = "Education")+  
 geom\_bar(mapping = aes(x = X\_bmi5cat, color = X\_educag), fill = NA)+  
 theme(axis.text.x = element\_text(size = 10, angle = 45, hjust = 1))



**Observation:** The segments concur with the correlation rawn above. Respondents with an education less than or equivalent to “Graduated from college or technical school”, tend to be less obses or over-weight

**Research quesion 2:**

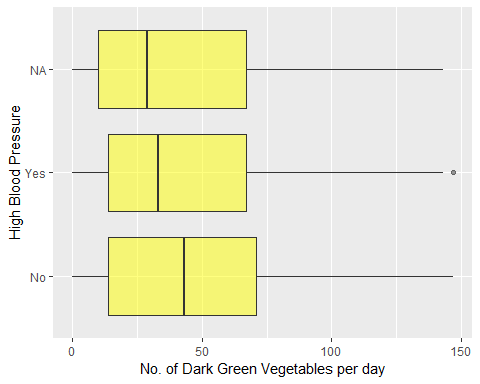
brfss2013 %>%  
 group\_by(X\_rfhype5) %>%  
 summarise(mean\_greenies = mean(grenday\_, na.rm = TRUE), mean\_bean = mean(beanday\_, na.rm = TRUE))

## Warning: Factor `X\_rfhype5` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

## # A tibble: 3 x 3  
## X\_rfhype5 mean\_greenies mean\_bean  
## <fct> <dbl> <dbl>  
## 1 No 58.1 28.7  
## 2 Yes 50.6 26.7  
## 3 <NA> 47.1 27.6

*Observation:* The results indicate that respondents who did not report having a high blood pressure tend to have a diet richer in green vegetables and beans.

brfss2013 %>%  
 filter(grenday\_ < 150) %>%  
 ggplot(aes(x = X\_rfhype5, y = grenday\_))+  
 labs(x = "High Blood Pressure", y = "No. of Dark Green Vegetables per day") +  
 geom\_boxplot(fill = "yellow", alpha = 0.5) +  
 coord\_flip()



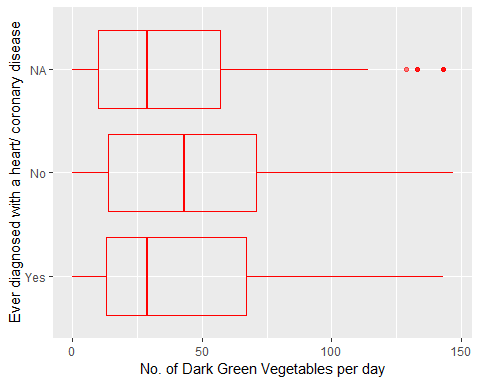
**Observation:** The boxplot clearly indicates that respondents with high blood pressure report to have lower median number of dark green vegetables consumed/day. The spread of the data under is the same for all 3 plots.

brfss2013 %>%  
 group\_by(cvdcrhd4) %>%  
 summarise(mean\_greenies = mean(grenday\_, na.rm = TRUE), mean\_bean = mean(beanday\_, na.rm = TRUE))

## Warning: Factor `cvdcrhd4` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

## # A tibble: 3 x 3  
## cvdcrhd4 mean\_greenies mean\_bean  
## <fct> <dbl> <dbl>  
## 1 Yes 47.2 27.1  
## 2 No 55.6 28.0  
## 3 <NA> 43.0 25.4

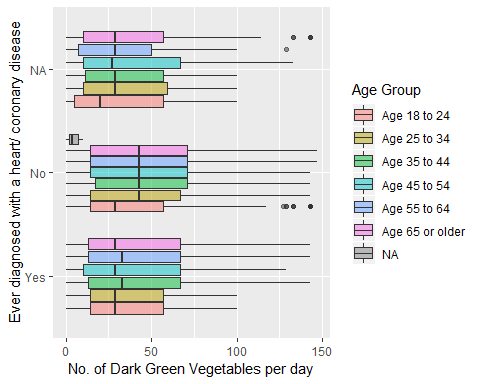
brfss2013 %>%  
 filter(grenday\_ < 150) %>%  
 ggplot(aes(x = cvdcrhd4, y = grenday\_))+  
 labs(x = "Ever diagnosed with a heart/ coronary disease", y = "No. of Dark Green Vegetables per day") +  
 geom\_boxplot(fill = NA, color = "red", alpha = 0.5) +  
 coord\_flip()



**Observation:** The boxplot suggests that respondents who have not been diagnosed with a coronary/ heart disease tend to have a diet richer in green vegetables (as shown by the median).

However, it is to be noted that the age may be a confounding variable in this case. Hence, we make another plot where we specifically take note of the age of the respondent (especially the elderly ones)

brfss2013 %>%  
 filter(grenday\_ < 150) %>%  
 ggplot(aes(x = cvdcrhd4, y = grenday\_, fill = X\_age\_g))+  
 labs(x = "Ever diagnosed with a heart/ coronary disease", y = "No. of Dark Green Vegetables per day", fill = "Age Group") +  
 geom\_boxplot(alpha = 0.5) +  
 coord\_flip()



**Observation:** It must be noted that the all respondents who are above the age 45 and have been diagnosed with heart/ coronary diseaeses tend to have a lower median consuumption of green vegetables.

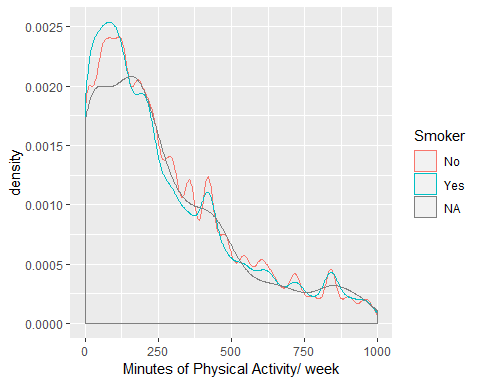
A causal inference can be made here: Lack of green vegetables is a cause of heart/ coronary ailments in the elderly respondents.

**Research quesion 3:**

brfss2013 %>%  
 summarise(mean\_phys = mean(pa1min\_, na.rm = TRUE), std\_phys = sd(pa1min\_, na.rm = TRUE))

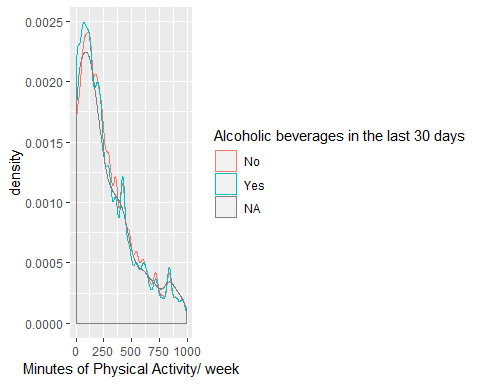
## mean\_phys std\_phys  
## 1 440.105 728.6152

p <- brfss2013 %>%  
 filter(pa1min\_ < 1000)%>%  
 ggplot(mapping = aes(x = pa1min\_))  
p + geom\_density(mapping = aes(color = X\_rfsmok3)) +  
 labs(x = "Minutes of Physical Activity/ week", color = "Smoker")



**Observation:** Respondents who do not smoke tend to have slightly longer hours of physical activity which indicates that smoking may not be just recreational and that smokers tend to be less cautious about physical fitness.

p + geom\_density(mapping = aes(color = drnkany5)) +   
 labs(x = "Minutes of Physical Activity/ week", color = "Alcoholic beverages in the last 30 days")

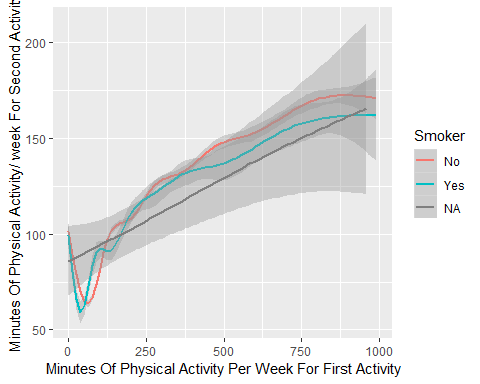


**Observation:** There is not a marked difference between the times of physical activity of those who have consumed drinks/ alcohol in the past 30 days. This result is confounding because the type of alcohol consumed and frequency of consumption are lurking variables.

brfss2013 %>%  
 filter(pamin11\_ < 1000 & pamin21\_ < 1000)%>%  
 ggplot(mapping = aes(x = pamin11\_, y = pamin21\_, ))+  
 geom\_smooth(mapping = aes(color = X\_rfsmok3, se = FALSE))+  
 labs(y = "Minutes Of Physical Activity/ week For Second Activity",   
 x = " Minutes Of Physical Activity Per Week For First Activity", color = "Smoker")

## Warning: Ignoring unknown aesthetics: se

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



**Observation:** There appears to be a strong positive correlation between the time a respondent devotes to two physical activities i.e. most respondents devote as much time to the first activity as they do to the second.