Exploring the BRFSS data

## Setup

### Load packages

library(ggplot2)  
library(dplyr)

Adding a few useful packages

library(ggthemes)  
library(scales)

### Load data

load("brfss2013.RData")  
attach(brfss2013)

## Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS) is a telephone survey that collects data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. It collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

**Population:** Health characteristics estimated from the BRFSS pertain to the non-institutionalized adult population, aged 18 years or older, who reside in the US.

Respondent data are forwarded to CDC to be aggregated for each state, returned with standard tabulations, and published at year’s end by each state. Source: <https://www.cdc.gov/brfss/>

**Random sampling?**

“The survey is conducted using Random Digit Dialing (RDD) techniques on both landlines and cell phones.” Source: <https://www.cdc.gov/brfss/about/brfss_faq.htm> Sampling appears to be random, therefore, results can be generalized to the population of the US.

*Bias:*

1.Selection bias:

There might be some selection bias introduced due to the survey being telephone-based. A proportion of the population of interest might not own telephones; landlines or cellular.

2.Response bias:

Also, response bias may arise from some people not answering their phones.

3.Recall bias:

Participants are asked questions to which they might not remember the answers.

**Random assignment?**

Since this is a cross-sectional study, it is observational, not interventional, So random assignment is not possible. Therefore, we cannot establish causality, only associations.

## Part 2: Research questions

**Research quesion 1:**

Is there an association between the duration of sleep and alcohol consumption? Given that sleep disorders are common in patients suffering from depression, does this relationship differ in those that have been diagnosed with a depressive disorder than those that haven’t?

Source: <https://www.sleepfoundation.org/articles/depression-and-sleep>

**Research quesion 2:**

Doctors constantly say that high cholesterol is bad for you and is a risk factor for heart disease. Does this data confirm that? Does the percentage of heart attack diagnosis increase with high cholesterol? And, since obesity is considered a risk factor for both high cholesterol and heart disease, is body mass index a confounder that affects the relationship between the two?

Sources: <https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/symptoms-causes/syc-20350800> <https://my.clevelandclinic.org/health/articles/17308-obesity--heart-disease>

**Research quesion 3:**

Does satisfaction with life increase with income? Since I’m currently broke, I’d like to know.

## Part 3: Exploratory data analysis

# Research quesion 1:

Difference between those with and without depressive disorder in mean sleeping duration and average alcohol drinking per day in the last 30 days.

**Variables of Interest**

1.*addepev2:* Ever Told You Had A Depressive Disorder

2.*sleptim1:* How Much Time Do You Sleep

3.*avedrnk2:* Avg Alcoholic Drinks Per Day In Past 30

First, getting to know the variables

class(addepev2)

## [1] "factor"

class(avedrnk2)

## [1] "integer"

class(sleptim1)

## [1] "integer"

Then, taking a quick look at the some of the observations

levels(addepev2)

## [1] "Yes" "No"

head(avedrnk2)

## [1] 2 NA 4 2 2 NA

head(sleptim1)

## [1] NA 6 9 8 6 8

Dealing with missing data by filtering NAs and rewriting our dataframe with the new “clean” one.

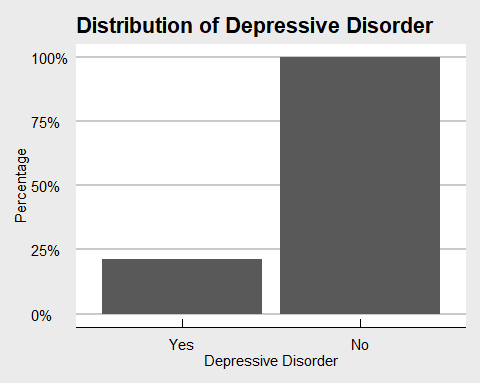
brfss2013 <- brfss2013 %>%  
 filter(!is.na(addepev2), !is.na(sleptim1), !is.na(avedrnk2))

**Visualizing the distribution of our variables**

**First,** the grouping variable:the percentage of people with depressive disorders

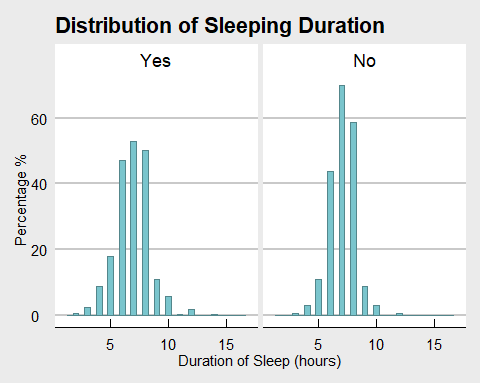
Geom\_bar will be used to create a bar plot since it is a categorical variable.

ggplot(data = brfss2013, aes(x = addepev2,y = stat(count/ max(count)))) +  
 geom\_bar(na.rm = T) +  
 scale\_y\_continuous(labels = percent\_format()) +  
 labs(x = "Depressive Disorder", y = "Percentage",  
 title = "Distribution of Depressive Disorder") +  
 theme\_economist\_white()



**Second,** The distribution of Sleeping Duration

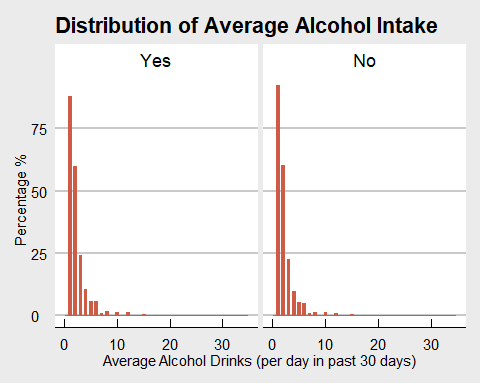
ggplot(data = brfss2013, aes(x = sleptim1, y = ..density..\*100)) +  
 geom\_histogram(color = "cadetblue4", fill = "cadetblue3", na.rm = T,binwidth = 0.5) +   
 xlim(1,17) +  
 labs(x = "Duration of Sleep (hours)", y = "Percentage %",  
 title = "Distribution of Sleeping Duration") +  
 facet\_wrap(~ addepev2) +  
 theme\_economist\_white()



The distribution appears to be nearly symmetrical in both groups, with a mean around 7 hours of sleep and very few outliers(a very small proportion of the population in both groups sleeps for more than 15 hours a day).

**Third,** The distribution of Average Alcohol Drinks

ggplot(data = brfss2013, aes(x = avedrnk2, y = ..density..\*100)) +  
 geom\_histogram(color = "coral3", fill = "coral3", na.rm = T, binwidth = 0.5) +  
 xlim(0,35) +  
 labs(x = "Average Alcohol Drinks (per day in past 30 days)", y = "Percentage %",  
 title = "Distribution of Average Alcohol Intake") +  
 facet\_grid(~ addepev2) +  
 theme\_economist\_white()



Unlike sleeping duration, the distribution of alcohol intake is heavily right skewed with a very notable outlier of 30 drinks per day.

**Performing a quick numerical summary**

Since the distribution of sleeping duration is symmetrical, we will use mean and standard deviation(SD) to summarize it. But median and Interquantile range(IQR) will be used to summarize the right-skewed distribution of average alcohol drinks.

brfss2013 %>%  
 group\_by(addepev2) %>%  
 summarise(count = n(), mean\_sleep = mean(sleptim1),   
 SD\_sleep = sd(sleptim1),   
 median\_drink = median(avedrnk2),   
 IQR\_drink = IQR(avedrnk2))

## # A tibble: 2 x 6  
## addepev2 count mean\_sleep SD\_sleep median\_drink IQR\_drink  
## <fct> <int> <dbl> <dbl> <int> <dbl>  
## 1 Yes 40179 6.92 1.60 2 2  
## 2 No 188739 7.06 1.20 2 1

*Summary of Sleep Duration:*

The mean daily sleep duration of those with depression is 6.9 hours with a SD of 1.6 hours, and the mean of those without depression is 7 hours with a SD of 1.2 hours. The difference between the two groups appears to be small with 95% of the population in both getting between approximately 5 and 8.5 hours of sleep dialy.

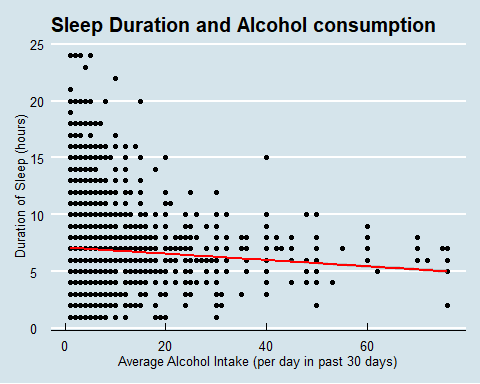
*Summary of Average Alcohol Drinks per day(in the last 30 days):*

The median average number of alcohol drinks per day of those with depression is 2 drinks with the middle 50% of the population(IQR) consuming between 0 and 4 drinks daily. Those without depression have the same median but with less variation as the IQR = 1(the middle 50% of the population consumes between 1 and 3 drinks daily).

**Visualing the relationship between the two variables**

First, the relationship between Sleep Duration and Alcohol consumption

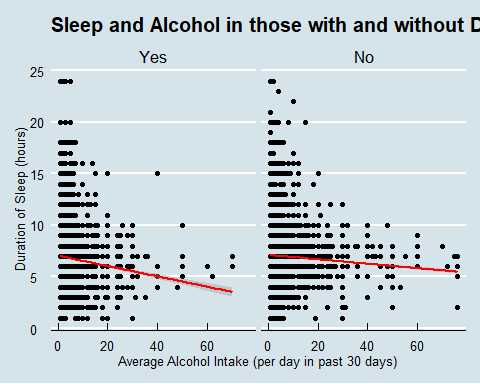
ggplot(data = brfss2013, aes(x = avedrnk2, y = sleptim1)) +  
 geom\_point() +   
 geom\_smooth(method = "lm", colour = "red") +  
 labs(x = "Average Alcohol Intake (per day in past 30 days)", y = "Duration of Sleep (hours)",  
 title = "Sleep Duration and Alcohol consumption") +  
 theme\_economist()



There does not seem to be a particular association- either positive or negative- between sleep duration and alcohol intake.

But, grouped by the presence of a depressive disorder

ggplot(data = brfss2013, aes(x = avedrnk2, y = sleptim1)) +  
 geom\_point(na.rm = TRUE) +   
 geom\_smooth(method = "lm", colour = "red") +  
 labs(x = "Average Alcohol Intake (per day in past 30 days)", y = "Duration of Sleep (hours)",  
 title = "Sleep and Alcohol in those with and without Depression") +  
 facet\_grid(~ addepev2) +  
 theme\_economist()



There appears to be an inverse association- albeit a small one- between sleep duration and alcohol intake. Large amounts of alcohol appear to slightly decrease sleep duaration,more so in those with depressive disorders. This association needs to be properly studied with regression models.

# Research quesion 2:

The relationship between high cholesterol and heart attacks among various categories of body mass index.

**Variables of Interest**

1.*X\_bmi5cat:* Computed Body Mass Index Categories

2.*cvdinfr4:* Ever Diagnosed With Heart Attack

3.*toldhi2:* Ever Told Blood Cholesterol High

First, getting to know the variables

class(X\_bmi5cat)

## [1] "factor"

class(cvdinfr4)

## [1] "factor"

class(toldhi2)

## [1] "factor"

Then, taking a quick look at some of the observations

levels(X\_bmi5cat)

## [1] "Underweight" "Normal weight" "Overweight" "Obese"

levels(cvdinfr4)

## [1] "Yes" "No"

levels(toldhi2)

## [1] "Yes" "No"

Dealing with missing data by filtering NAs and rewriting our dataframe with the new “clean” one.

brfss2013 <- brfss2013 %>%  
 filter(!is.na(X\_bmi5cat), !is.na(cvdinfr4), !is.na(toldhi2))

**Summarizing and visualizing the distribution of our variables**

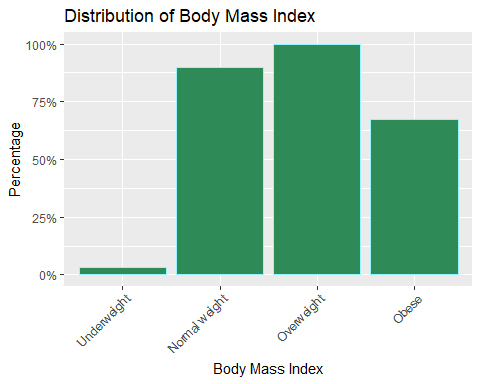
Geom\_bar will be used to create bar plots since they are categorical variables.

**First,** body mass index

table(X\_bmi5cat)

## X\_bmi5cat  
## Underweight Normal weight Overweight Obese   
## 8267 154898 167084 134799

ggplot(data = brfss2013, aes(x = X\_bmi5cat,y = stat(count/ max(count)))) +  
 geom\_bar(color = "cadetblue1", fill = "seagreen4") +  
 scale\_y\_continuous(labels = percent\_format()) +  
 labs(x = "Body Mass Index", y = "Percentage") +  
 ggtitle("Distribution of Body Mass Index") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



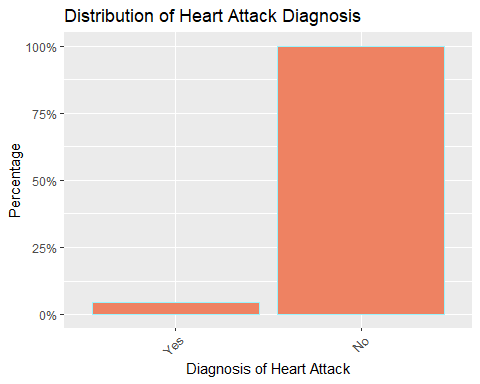
The plot shows that the majority of the population is either overweight or obese, as confirmed by the summary.

**Second,** diagnosis of heart attack

table(cvdinfr4)

## cvdinfr4  
## Yes No   
## 29284 459904

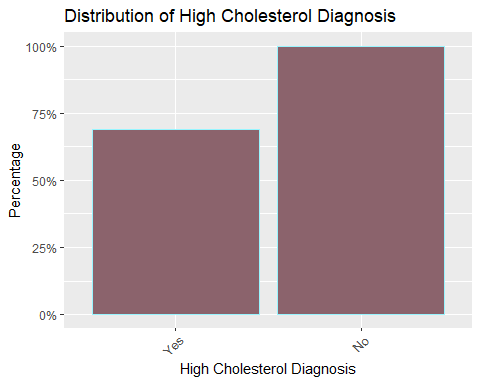
ggplot(data = brfss2013, aes(x = cvdinfr4,y = stat(count/ max(count)))) +  
 geom\_bar(color = "cadetblue1", fill = "salmon2") +  
 scale\_y\_continuous(labels = percent\_format()) +  
 labs(x = "Diagnosis of Heart Attack", y = "Percentage") +  
 ggtitle("Distribution of Heart Attack Diagnosis") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



The majority of the population were not diagnosed with heart attacks.

**Third,** diagnosis of high cholesterol

ggplot(data = brfss2013, aes(x = toldhi2,y = stat(count/ max(count)))) +  
 geom\_bar(color = "cadetblue1", fill = "pink4") +  
 scale\_y\_continuous(labels = percent\_format()) +  
 labs(x = "High Cholesterol Diagnosis", y = "Percentage") +  
 ggtitle("Distribution of High Cholesterol Diagnosis") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



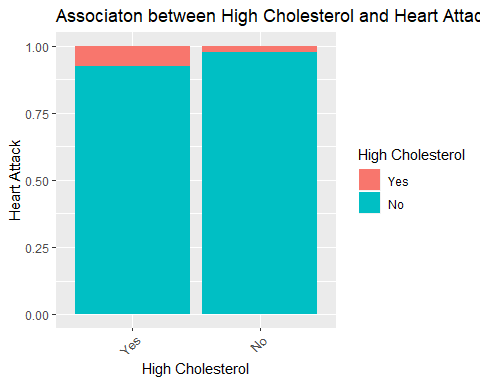
The percentage of high cholesterol seems to be relatively very high, which is expected, given the high percentage of obesity in our sample.

**Summarizing and visualizing the relationship between high cholesterol and heart attack diagnosis**

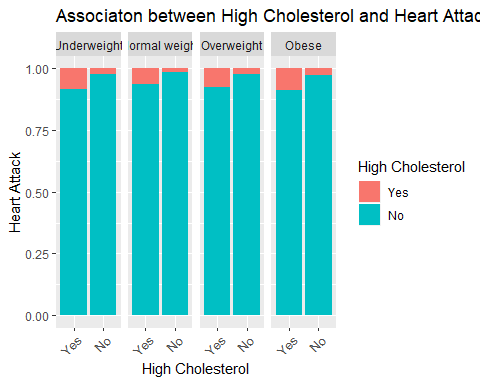
table(cvdinfr4, toldhi2)

## toldhi2  
## cvdinfr4 Yes No  
## Yes 19560 8299  
## No 162778 227472

ggplot(brfss2013, aes(x = toldhi2, fill = cvdinfr4)) +  
 geom\_bar(position = "fill",) +  
 labs(x = "High Cholesterol", y = "Heart Attack") +  
 ggtitle("Associaton between High Cholesterol and Heart Attacks") +  
 scale\_fill\_discrete(name = "High Cholesterol") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



ggplot(brfss2013, aes(x = toldhi2, fill = cvdinfr4)) +  
 geom\_bar(position = "fill",) +  
 labs(x = "High Cholesterol", y = "Heart Attack") +  
 ggtitle("Associaton between High Cholesterol and Heart Attacks") +  
 scale\_fill\_discrete(name = "High Cholesterol") +  
 facet\_grid(~ X\_bmi5cat) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



The plots and the summary show a positive association between high cholesterol and the percentage of heart attack diagnosis. More people that have high cholesterol have also been diagnosed with heart attacks than those with low cholesterol. The association between the two variables does not seem to change with various categories of body mass index. But it is yet to be confirmed and quantified with statistical tests and models.

**Research quesion 3:**

The relationship between income and satisfation with life

**Variables of Interest**

1.*income2:* Income Level

2.*lsatisfy:* Satisfaction With Life

First, getting to know the variables

class(income2)

## [1] "factor"

class(lsatisfy)

## [1] "factor"

Then, taking a quick look at some of the observations

levels(income2)

## [1] "Less than $10,000" "Less than $15,000" "Less than $20,000"  
## [4] "Less than $25,000" "Less than $35,000" "Less than $50,000"  
## [7] "Less than $75,000" "$75,000 or more"

levels(lsatisfy)

## [1] "Very satisfied" "Satisfied" "Dissatisfied"   
## [4] "Very dissatisfied"

Dealing with missing data by filtering NAs and rewriting our dataframe with the new “clean” one.

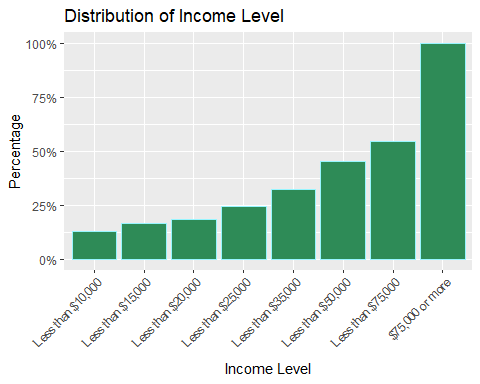
brfss2013 <- brfss2013 %>%  
 filter(!is.na(income2), !is.na(lsatisfy))

**Visualizing the distribution of our variables**

Geom\_bar will be used to create bar plots since they are categorical variables.

**First,** income level

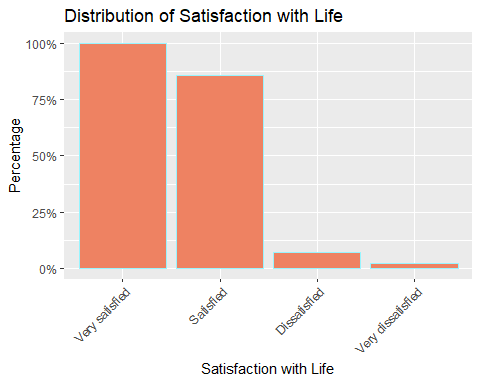
ggplot(data = brfss2013, aes(x = income2,y = stat(count/ max(count)))) +  
 geom\_bar(color = "cadetblue1", fill = "seagreen4") +  
 scale\_y\_continuous(labels = percent\_format()) +  
 labs(x = "Income Level", y = "Percentage") +  
 ggtitle("Distribution of Income Level") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



The distribution of income level is surprisingly left-skewed, with 50% of the population earning more than 50,000$. Since, it is usually expected to be right-skewed- as less people tend to earn that large amount of money- this suggests that the sample is not representative of the true population.

**Second,** Satisfaction with life

ggplot(data = brfss2013, aes(x = lsatisfy, y = stat(count/ max(count)))) +  
 geom\_bar(color = "cadetblue1", fill = "salmon2") +  
 scale\_y\_continuous(labels = percent\_format()) +  
 labs(x = "Satisfaction with Life", y = "Percentage") +  
 ggtitle("Distribution of Satisfaction with Life") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



The plot shows that the majority of the population are satisfied or very satisfied with life, with less than 15% dissatisfied or very dissatisfied.

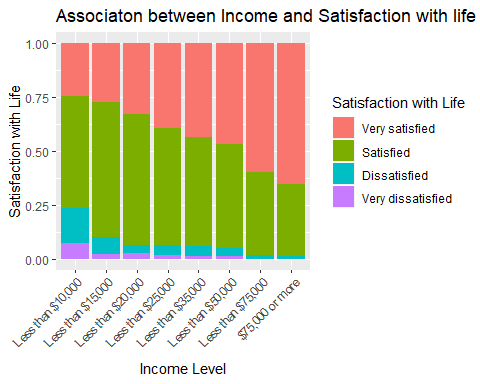
**Performing a quick numerical summary**

table(income2, lsatisfy)

## lsatisfy  
## income2 Very satisfied Satisfied Dissatisfied  
## Less than $10,000 238 461 113  
## Less than $15,000 309 548 101  
## Less than $20,000 374 594 52  
## Less than $25,000 398 572 58  
## Less than $35,000 517 587 67  
## Less than $50,000 634 589 43  
## Less than $75,000 708 504 28  
## $75,000 or more 1112 563 28  
## lsatisfy  
## income2 Very dissatisfied  
## Less than $10,000 52  
## Less than $15,000 27  
## Less than $20,000 19  
## Less than $25,000 16  
## Less than $35,000 8  
## Less than $50,000 8  
## Less than $75,000 1  
## $75,000 or more 3

**Visualizing the relationship between the two variables**

ggplot(brfss2013, aes(x = income2, fill = lsatisfy)) +  
 geom\_bar(position = "fill",) +  
 labs(x = "Income Level", y = "Satisfaction with Life") +  
 ggtitle("Associaton between Income and Satisfaction with life") +  
 scale\_fill\_discrete(name = "Satisfaction with Life") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



Both the table and the plot appear to show that there is a positive relationship between the level of income and satisfation with life. Those with higher income tend to be more satisfied with their lives. This finding needs to be confirmed and quantified with regression models and statistical tests.