PROJECT\_STATISTIC

NGUYEN NGOC PHUNG

12/23/2020

## INTRODUCTION

This project almost places focus on data of BRFSS in 2013 by analyzing, visualizing and using techniques (ex: hypothesis testing) in order to answer some questions. The Behavioral Risk Factor Surveillance System (BRFSS) is a collaborative project between all of the states in the United States (US) and participating US territories and the Centers for Disease Control and Prevention (CDC). The BRFSS is administered and supported by CDC’s Population Health Surveillance Branch, under the Division of Population Health at the National Center for Chronic Disease Prevention and Health Promotion. BRFSS is an ongoing surveillance system designed to measure behavioral risk factors for the non-institutionalized adult population (18 years of age and older) residing in the US.

The brief summary of BRFSS data is described in the following HTML file (Note:you have to download to open the file)

[BRFSS data](https://drive.google.com/file/d/1Cun2EU544Muhw9JoETFYy3yEAStEG9Y1/view?usp=sharing)

The data includes the following:

MAIN SURVEY: Main survey is divided into 17 Sections

Section 0 - Record Identification

Section 1 - Health Status

Section 2 - Healthy Days - Health-Related Quality of Life

Section 3 - Health Care Access

Section 4 - Inadequate Sleep

Section 6 - Cholesterol Awareness

Section 7 - Chronic Health Conditions

Section 8 - Demographics

Section 9 - Tobacco Use

Section 10 - Alcohol Consumption

Section 11 - Fruits and Vegetables

Section 12 - Exercise (Physical Activity)

Section 13 - Arthritis Burden

Section 14 - Seatbelt Use

Section 15 - Immunization

Section 16 - HIV/AIDS

OPTIONAL MODULE: Optional module contains 22 small modules. However, we will not use data from this part.

## DATA AND RESULTS

### LOAD PACKAGE AND DATA

# LOAD PACKAGE AND DATA @-@  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(cowplot)

## Warning: package 'cowplot' was built under R version 4.0.3

library(plotly)

## Warning: package 'plotly' was built under R version 4.0.3

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

load('brfss2013.RData')

### QUESTION 1 (PIE CHART)

#### COMPARE THE GENERAL HEALTH OF MALE AND FEMALE

#### DATA STRUCTURE

# This question we use 2 main attributes:  
# genhlth: General Health   
# sex: respondents sex  
str(brfss2013$genhlth)

## Factor w/ 5 levels "Excellent","Very good",..: 4 3 3 2 3 2 4 3 1 3 ...

str(brfss2013$sex)

## Factor w/ 2 levels "Male","Female": 2 2 2 2 1 2 2 2 1 2 ...

#### STEP 1: (FILTER DATA) GET THE DATA FRAME OF FEMALE AND MALE

female\_Q1<-brfss2013 %>%   
 filter(as.integer(sex)==2,!is.na(genhlth))  
  
male\_Q1<-brfss2013 %>%   
 filter(as.integer(sex)==1,!is.na(genhlth))

#### STEP 2: PLOT 2 PIE CHARTS OF FEMALE AND MALE WITH RESPECT TO THE GENERAL HEALTH

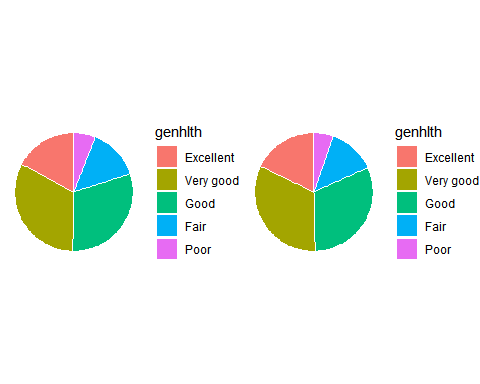
sum\_F<-female\_Q1 %>%  
 count(genhlth)  
  
sum\_F

## genhlth n  
## 1 Excellent 49740  
## 2 Very good 93940  
## 3 Good 87557  
## 4 Fair 40844  
## 5 Poor 17238

sum\_M<-male\_Q1 %>%  
 count(genhlth)  
  
sum\_M

## genhlth n  
## 1 Excellent 35741  
## 2 Very good 65135  
## 3 Good 62998  
## 4 Fair 25882  
## 5 Poor 10713

plot\_F<-ggplot(sum\_F, aes(x = "", y = n, fill = genhlth)) +  
 geom\_bar(width = 1, stat = "identity", color = "white") +  
 coord\_polar("y", start = 0)+  
 theme\_void()  
  
plot\_M<-ggplot(sum\_M, aes(x = "", y = n, fill = genhlth)) +  
 geom\_bar(width = 1, stat = "identity", color = "white") +  
 coord\_polar("y", start = 0)+  
 theme\_void()  
  
plot\_grid(plot\_F,plot\_M)



#### CONCLUSION

#### By the charts we can see that proportion of general health of male and female is merely the same, with categories “very good” and “good” hold the majority *-*

### QUESTION 2: (SEGMENTED BAR CHART)

#### IS EDUCATION LEVEL CORRELATED WITH INCOME LEVEL AND EMPLOYMENT STATUS ?

#### DATA STRUCTURE

# INTRO  
# This question we use 3 main attributes:  
# educa: Enducation level  
# income2: Income level  
# employ1: Employment Status  
  
str(brfss2013$educa)

## Factor w/ 6 levels "Never attended school or only kindergarten",..: 6 5 6 4 6 6 4 5 6 4 ...

str(brfss2013$income2)

## Factor w/ 8 levels "Less than $10,000",..: 7 8 8 7 6 8 NA 6 8 4 ...

str(brfss2013$employ1)

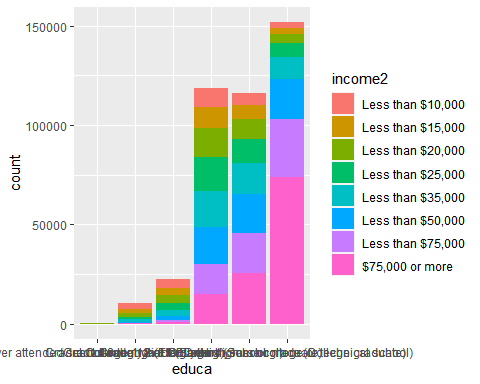
## Factor w/ 8 levels "Employed for wages",..: 7 1 1 7 7 1 1 7 7 5 ...

#### STEP 1: (FILTER DATA) GET THE DATA FRAME FOR INCOME LEVEL AND EMPLOYMENT STATUS

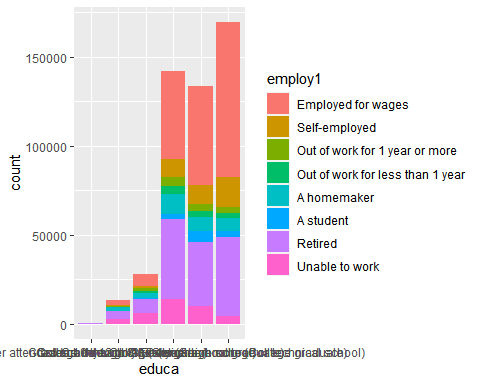
income\_Q2<-brfss2013 %>%  
 filter(!is.na(educa),!is.na(income2))  
  
employ\_Q2<-brfss2013 %>%  
 filter(!is.na(educa),!is.na(employ1))

#### STEP 2: PLOT 2 SEGMENTED BAR CHARTS TO SEE THE RELATIONSHIP BETWEEN (EDUCATION & INCOME) AND (EDUCATION & EMPLOYMENT)

ggplot(data = income\_Q2, aes(x = educa, fill = income2)) +  
 geom\_bar()



ggplot(data = employ\_Q2, aes(x = educa, fill = employ1)) +  
 geom\_bar()



#### CONCLUSION

#### (Education & Income)

#### The bar chart show that education level and income level have strong correlation, which means that higher education level will lead to higher salary. This is best demonstrated in the 2 last column, people who have education level at college 4 years or more have the income from $75,000 or more 3 times higher than people who have education level at 1 year to 3 years.

#### (Education & Employment)

#### The second bar chart show the correlation between education level and employment status. We can conclude from the chart that people who have education level at grade 12 or higher, will have much higher chance to be employed compare to the others.

### QUESTION 3: (BOX PLOT)

#### WHAT IS THE DIFFERENCES IN HOURS OF SLEEP BETWEEN THOSE DO HAVE DIFFICULTY CONCENTRATING AND REMEMBERING AND THOSE WHO DO NOT @-@

#### DATA STRUCTURE

# This question we use 2 main attributes:  
# sleptim1:How much time do you sleep   
# decide: Difficulty concentrating or remembering  
str(brfss2013$sleptim1)

## int [1:491775] NA 6 9 8 6 8 7 6 8 8 ...

str(brfss2013$decide)

## Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...

#### STEP 1: (FILTER DATA) GET THE DATA FRAME FOR # OF SLEEP AND DIFFICULTY CONCENTRATING AND REMEMBERING

Q3<-brfss2013 %>%  
 filter(!is.na(sleptim1),!is.na(decide))

#### STEP 2: SUMMARIZE HOURS OF SLEEP GROUP BY PEOPLE HAVING CONCENTRATING AND REMEMBERING OR NOT @-@

Q3 %>%   
 group\_by(decide)%>%  
 summarise(mean=mean(sleptim1), median=median(sleptim1),Var=var(sleptim1),IQR=IQR(sleptim1),n=n())

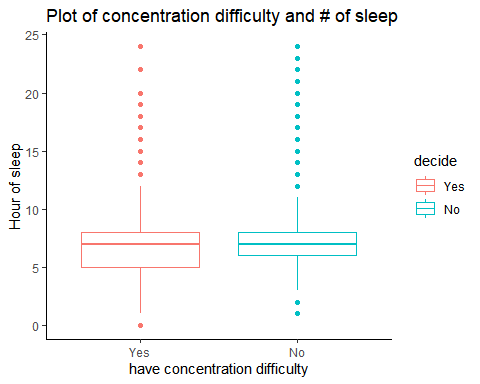
## `summarise()` ungrouping output (override with `.groups` argument)

## Warning: `...` is not empty.  
##   
## We detected these problematic arguments:  
## \* `needs\_dots`  
##   
## These dots only exist to allow future extensions and should be empty.  
## Did you misspecify an argument?

## # A tibble: 2 x 6  
## decide mean median Var IQR n  
## <fct> <dbl> <dbl> <dbl> <dbl> <int>  
## 1 Yes 6.75 7 4.67 3 48860  
## 2 No 7.09 7 1.83 2 423099

#### STEP 3: DRAW THE BOX PLOT

ggplot(Q3, aes(x=decide, y=sleptim1, color=decide)) +  
 geom\_boxplot()+  
 labs(title="Plot of concentration difficulty and # of sleep",x="have concentration difficulty", y = "Hour of sleep")+  
 theme\_classic()



#### CONCLUSION

#### Those who have concentration difficulty have their daily number of sleeping hours vary more than those do not have

#### The mean hours of sleep of those having concentration difficulty is slightly smaller than the other (6.75<7.09)

### QUESTION 4: (GEXAGONAL HEATMAP)

#### HOW MUCH SHOULD WE SLEEP IN ORDER TO MAINTAIN A GOOD PHYSICAL AND MENTAL HEALTH ? @-@

#### SIMILARLY, HOW MUCH SHOULD WE SLEEP IN ORDER TO MAINTAIN A GOOD GENERAL HEALTH? DOES THE RESULT CONSISTENT WITH ABOVE RESULT? @-@

#### DATA STRUCTURE

# This question we use 4 main attributes:  
# sleptim1:How much time do you sleep   
# physhlth:Number Of Days Physical Health Not Good  
# menthlth:Number Of Days Mental Health Not Good  
# genhlth: General Health  
  
str(brfss2013$sleptim1)

## int [1:491775] NA 6 9 8 6 8 7 6 8 8 ...

str(brfss2013$physhlth)

## int [1:491775] 30 0 3 2 10 0 1 5 0 0 ...

str(brfss2013$menthlth)

## int [1:491775] 29 0 2 0 2 0 15 0 0 0 ...

str(brfss2013$genhlth)

## Factor w/ 5 levels "Excellent","Very good",..: 4 3 3 2 3 2 4 3 1 3 ...

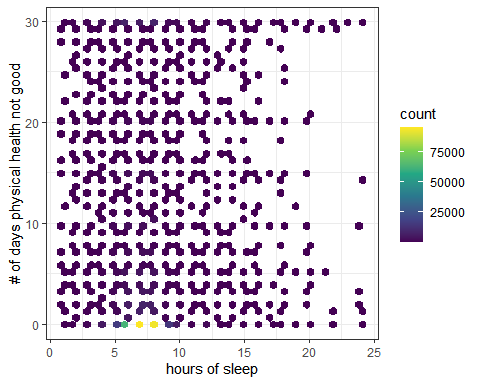
#### STEP 1: (FILTER DATA) GET THE DATA FRAME FOR NUMBER OF DAYS PHYSICAL HEALTH AND MENTAL HEALTH NOT GOOD @-@

#### Note: Since we have some outliers, hence, we have to remove all of them.

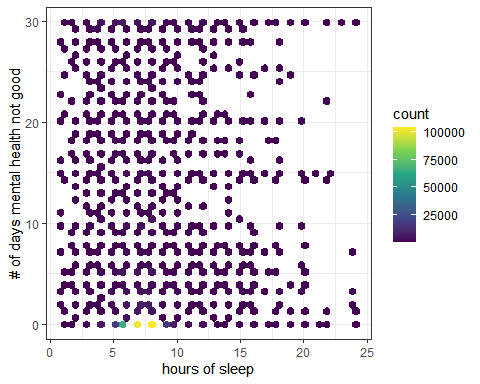
health<-brfss2013 %>%  
 filter(!is.na(sleptim1),!is.na(physhlth),!is.na(menthlth),!is.na(genhlth),sleptim1<=24,menthlth<=31)

#### STEP 2: DRAW THE GEXAGONAL HEATMAP FOR BOTH PHYSICAL AND MENTAL HEALTH AND COMPARE IT TO GENERAL HEALTH

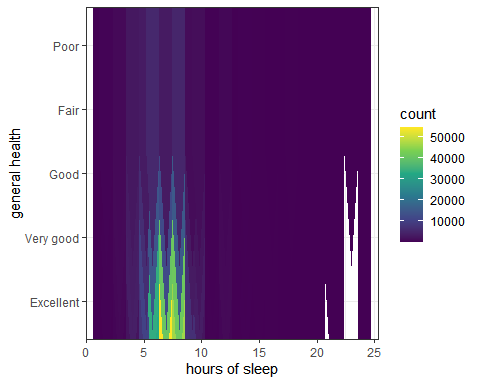
# Physical health  
ggplot(health,aes(sleptim1,physhlth))+  
 geom\_hex(bins=40)+  
 labs(colour="count", x="hours of sleep", y="# of days physical health not good")+  
 scale\_fill\_continuous(type="viridis") +   
 theme\_bw()



# Mental health  
ggplot(health,aes(sleptim1,menthlth))+  
 geom\_hex(bins=40)+  
 labs(colour="count", x="hours of sleep", y="# of days mental health not good")+  
 scale\_fill\_continuous(type="viridis") +   
 theme\_bw()



# General health  
ggplot(health,aes(sleptim1,genhlth)) +   
 geom\_hex(bins=40)+   
 labs(colour="count", x="hours of sleep", y="general health")+  
 scale\_fill\_continuous(type="viridis") +  
 theme\_bw()



#### CONCLUSION:

#### PHYSICAL &MENTAL: From the heatmap of physical and mental health, we see that those who sleep with recommended 6-9 hours a day have smaller number of days having physical and mental health in poor condition about 0-10 days. Moreover, if people sleep with ideal hours, the number of people with many days of having physical and mental health in poor condition will decrease.

#### GENERAL: Clearly, we see that the above result is consistent with the heatmap for general health with those who sleep 6-9 hours a day told that they have very good or excellent general health. Thus, people who sleep over 9 hours and have large number days of having physical and mental health in poor condition usually have fair and poor gereral health.

### QUESTION 5: (HYPOTHESIS TESTING + DENSITY PLOT)

#### Let X: weight of male respondents of population

#### Y: weight of male respondents of population

#### Assumption: X ~ N(u1,var1) , Y ~ N(u2,var2)

#### Check the following hypothesis with 95% confidence interval

#### H0: u1 - u2 = 37 (pounds)

#### Ha: u1 - u2 != 37 (pounds)

#### DATA STRUCTURE

# This question we use 2 main attributes:  
# sex: Respondents Sex  
# weight2: Reported Weight In Pounds  
str(brfss2013$sex)

## Factor w/ 2 levels "Male","Female": 2 2 2 2 1 2 2 2 1 2 ...

str(brfss2013$weight2)

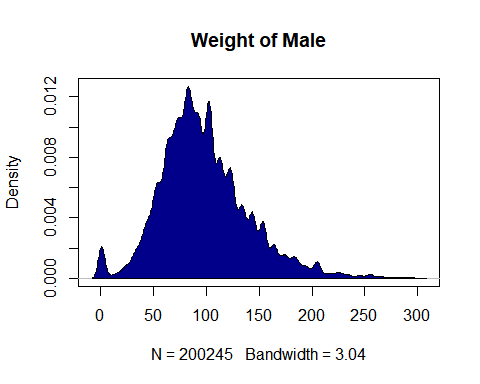
## Factor w/ 570 levels "",".b","100",..: 154 30 63 31 169 128 9 1 139 73 ...

#### STEP 1: (FILTER DATA) GET THE DATA FRAME FOR WEIGHT OF MALE AND FEMALE

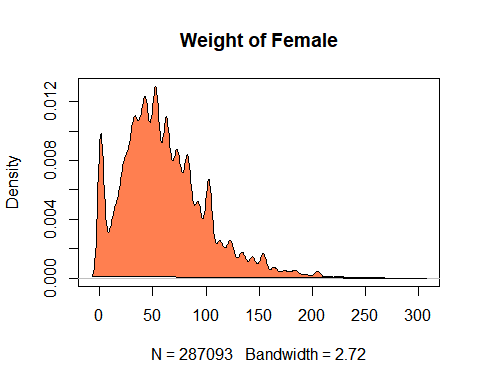
Male<-brfss2013 %>%  
 filter(!is.na(weight2),!is.na(sex),as.integer(weight2)<300,as.integer(sex)==1)  
  
Female<-brfss2013 %>%  
 filter(!is.na(weight2),!is.na(sex),as.integer(weight2)<300,as.integer(sex)==2)  
  
Male\_weight<-as.integer(Male$weight2)  
Female\_weight<-as.integer(Female$weight2)

#### STEP 2: VISUALIZE THE DATA FOR MALE AND FEMALE BY DENSITY PLOT

M <- density(Male\_weight)  
plot(M, main="Weight of Male")  
polygon(M, col="blue4", border="black")



# so # of observation for male is 200245  
n<-200245  
  
F <- density(Female\_weight)  
plot(F, main="Weight of Female")  
polygon(F, col="coral", border="black")



# so # of observation for female is 287093  
m<-287093

#### STEP 3: CALCULATE TEST STATISTIC

#### Denote V(X) and V(Y) be the sample variance of (X1,…,Xn) and (Y1,…,Ym) respectively

#### Since both m,n are large , we have TS=(mean(X)-mean(Y)-30)/sqrt(V(X)/n + V(Y)/m)

mean\_X<-mean(Male\_weight)  
mean\_Y<-mean(Female\_weight)  
V\_X<- (sd(Male\_weight))^2  
V\_Y<- (sd(Female\_weight))^2  
P<-matrix(c(mean\_X,V\_X,mean\_Y,V\_Y),2,2)  
dimnames(P)<-list(c("mean","sample var"),c("X","Y"))  
P

## X Y  
## mean 98.17198 61.52105  
## sample var 1781.53227 1697.22858

Test\_stat<-(mean\_X-mean\_Y-37)/sqrt((V\_X/n)+(V\_Y/m))  
Test\_stat

## [1] -2.868522

#### STEP 4: CALCULATE P-VALUE

#### we have p-value= 2P(Z>abs(Test statistic))

1- pnorm(abs(Test\_stat))

## [1] 0.002061974

#### CONCLUSION:

#### Since anpha= 5% > p-value= 0.002061974. Hence, we reject the null-hypothesis H0

#### For any anpha >= 3%, we have appropriate evidence to reject the null-hypothesis

#### For any anpha <= 2%, we fail to reject the null-hypothesis

#### From the plot of weight of male, we can see that men who have weight from 50 to 150 pounds have large density. Moreover, men with weight approximate 100 pounds have large density to 0.012.

#### From the plot of weight of female, we can see that women who have weight from 50 to 100 pounds have large density. Moreover, women with weight about 50 pounds have large density to 0.012.

#### Thus, at the 5% significance level, there is inufficient evidence to support the claim that the mean score on the test was equal 37 pounds.

### QUESTION 6: (3D PLOT)

#### Response variable: (poorhlth) number of days having poor physical or mental health in the last 30 days

#### Predictor: Physical exercise, diet (fruit and vegetables), hours of sleep

#### DATA STRUCTURE

str(brfss2013$exerany2)

## Factor w/ 2 levels "Yes","No": 2 1 2 1 2 1 1 1 1 1 ...

str(brfss2013$poorhlth)

## int [1:491775] 30 NA 0 0 0 NA 0 10 NA NA ...

str(brfss2013$fruit1)

## int [1:491775] 104 301 203 306 302 206 325 320 101 202 ...

str(brfss2013$sleptim1)

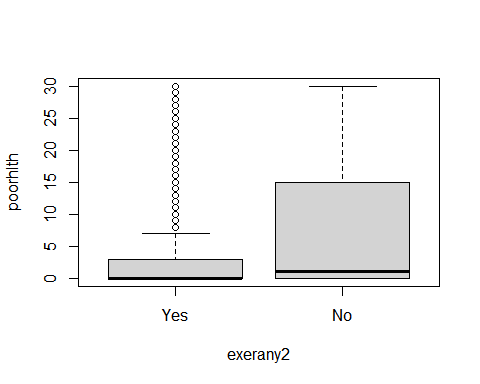
## int [1:491775] NA 6 9 8 6 8 7 6 8 8 ...

#### STEP 1:(FILTER DATA) GET DATA FRAME

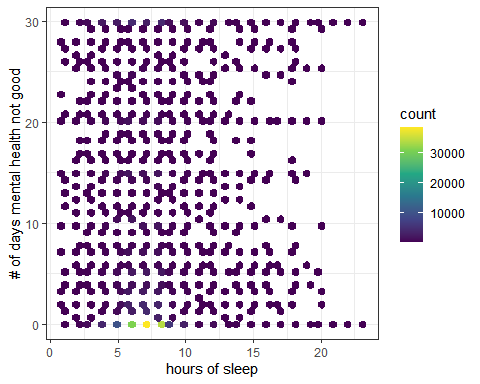
data<-brfss2013 %>%  
 filter(!is.na(exerany2),!is.na(poorhlth),!is.na(fruit1),!is.na(sleptim1),poorhlth<=30,sleptim1<24,fruit1<400)

#### STEP 2: PLOT DATA (3D PLOT)

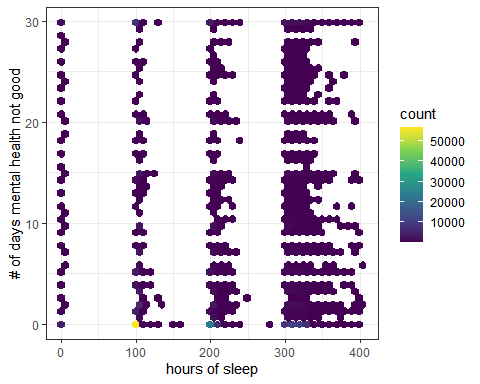
boxplot( poorhlth~exerany2, data = data)

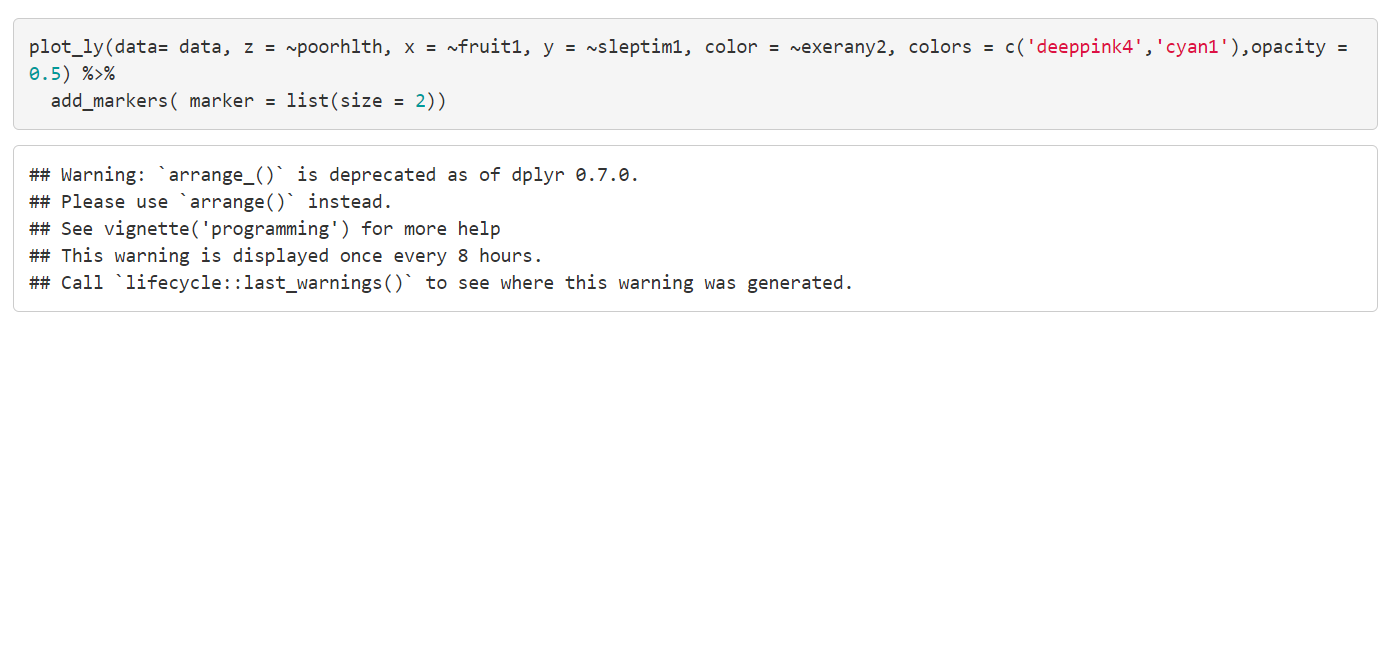
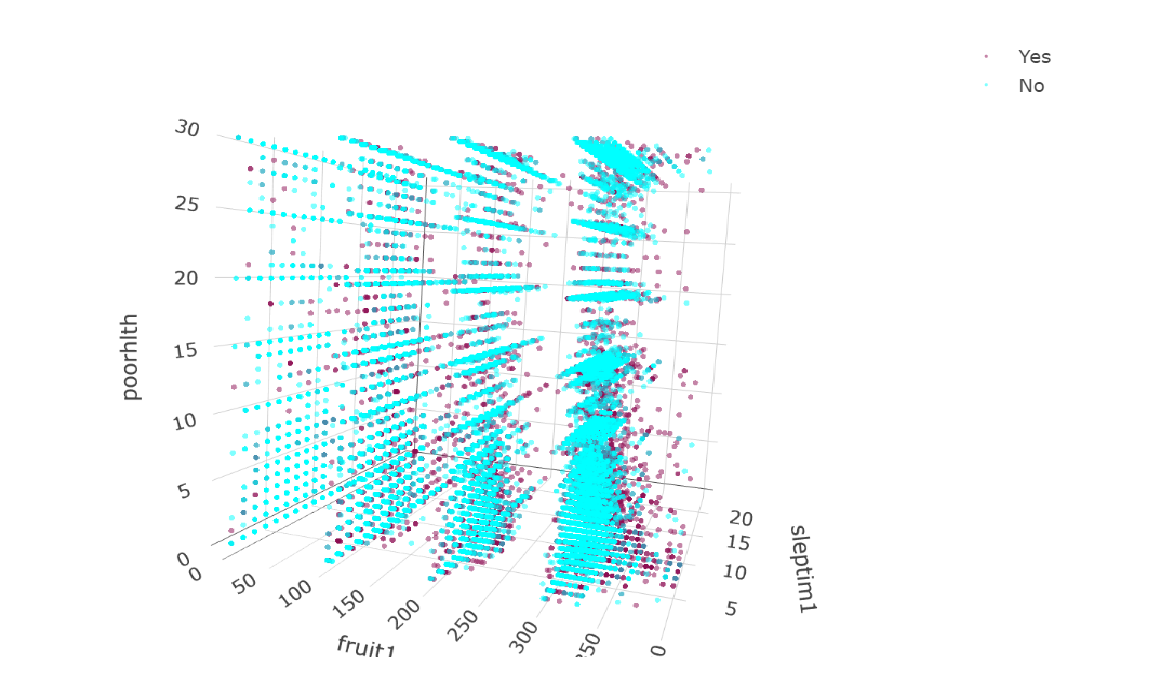


ggplot(data,aes(sleptim1,poorhlth))+  
 geom\_hex(bins=40)+  
 labs(colour="count", x="hours of sleep", y="# of days mental health not good")+  
 scale\_fill\_continuous(type="viridis") +   
 theme\_bw()



ggplot(data,aes(fruit1,poorhlth))+  
 geom\_hex(bins=40)+  
 labs(colour="count", x="hours of sleep", y="# of days mental health not good")+  
 scale\_fill\_continuous(type="viridis") +   
 theme\_bw()



[]( 

#### CONCLUSION

#### Those respondents who usually do exercise, eat fruit and vegetables and get adequate sleep are less likely to have poor physical and mental health than others.

#### Generally, physical exercise, diet and sleep quality all play important role in maintaining general good health.

## REFERRENCE

[BRFSS web site](http://www.cdc.gov/brfss/)

[BRFSS Questionnaire (Mandatory and Optional Modules)](http://www.cdc.gov/brfss/questionnaires/pdf-ques/2013%20BRFSS_English.pdf)

[BRFSS Codebook](http://www.cdc.gov/brfss/annual_data/2013/pdf/CODEBOOK13_LLCP.pdf)

[BRFSS Guide to Calculated Variables](http://www.cdc.gov/brfss/annual_data/2013/pdf/2013_Calculated_Variables_Version15.pdf)

[BRFSS Guide to Optional Modules Used, by State](http://apps.nccd.cdc.gov/BRFSSModules/ModByState.asp?Yr=2013)