Exploring the BRFSS data

Pranjal Vijay

December 12, 2017

Setup

Load packages

```
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
```

Load data

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

Part 1: Data

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). 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. In this document, the term “state” is used to refer to all areas participating in BRFSS, including the District of Columbia, Guam, and the Commonwealth of Puerto Rico. The BRFSS objective is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population. Factors assessed by the BRFSS in 2013 include tobacco use, HIV/AIDS knowledge and prevention, exercise, immunization, health status, healthy days - health-related quality of life, health care access, inadequate sleep, hypertension awareness, cholesterol awareness, chronic health conditions, alcohol consumption, fruits and vegetables consumption, arthritis burden, and seatbelt use. Since 2011, BRFSS conducts both landline telephone- and cellular telephone-based surveys. In conducting the BRFSS landline telephone survey, interviewers collect data from a randomly selected adult in a household. In conducting the cellular telephone version of the BRFSS questionnaire, interviewers collect data from an adult who participates by using a cellular telephone and resides in a private residence or college housing. The full BRFSS data set has been reduced to a selected group of variables, making 491775 observations of 336 variables.

I want to take a quick look of below data frame & viewing its dimensions.

```
brfss2013 %>%
  select(medcost, genhlth) %>%
  str()
```

```
## 'data.frame':    491775 obs. of  2 variables:
##  $ medcost: Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...
##  $ genhlth: Factor w/ 5 levels "Excellent","Very good",..: 4 3 3 2 3 2 4 3 1 3 ...
```

it can be seen that this data frame of 491775 obs. of 2 variables

I want to take a quick look of below data frame & viewing its dimensions.

```
brfss2013 %>%
  select(sleptim1, qlhlth2) %>%
  str()
```

```
## 'data.frame':    491775 obs. of  2 variables:
##  $ sleptim1: int  NA 6 9 8 6 8 7 6 8 8 ...
##  $ qlhlth2 : int  0 25 2 20 NA NA NA NA NA NA ...
```

it can be seen that this data frame of 491775 obs. of 2 variables

I want to take a quick look of below data frame & viewing its dimensions.

```
brfss2013 %>%
  select(painact2, exerany2) %>%
  str()
```

```
## 'data.frame':    491775 obs. of  2 variables:
##  $ painact2: int  5 0 20 0 NA NA NA NA NA NA ...
##  $ exerany2: Factor w/ 2 levels "Yes","No": 2 1 2 1 2 1 1 1 1 1 ...
```

it can be seen that this data frame of 491775 obs. of 2 variables * * *

Part 2: Research questions

Research quesion 1:

As mentioned, BRFSS data set is designed to measure behavioral risk factors for adults population. I want to find out whether the peroson who could not see the doctor due to cost have good health or not?. I need to analyse whether Cost affects general health for a person or not, if yes then how much amount of person affected. Therefore I have included two variables, first is medcost and second is genhlth. * * *

Research quesion 2:

I also want to find out whether the peroson who has good sleep also has healthy days or i.e How Many Days Full Of Energy In Past 30 Days? I need to analyse whether good sleep impacts on healthy days or not with graphical visualization? Therefore I have included two variables, first is sleptim1 and second is qlhlth2. * * *

Research quesion 3:

I want to find out whether the peroson who did not exercise in last 30 days also feels hard to do Usual activities in past 30 days or not? I need to analyse whether Not doing exercise makes hard to do usual activities or not with graphical visualization? Therefore I have included two variables, first is exerany2 and second is painact2. * * *

Part 3: Exploratory data analysis

first, I want to do some exploratory data analysis to find out solutions of above questions.

Research quesion 1:

Solution start from below

Here tabulating general_health data is being done

```
brfss2013 %>%
  group_by(genhlth) %>%
  summarise(count = n())
```

```
## # A tibble: 6 x 2
##     genhlth  count
##      <fctr>  <int>
## 1 Excellent  85482
## 2 Very good 159076
## 3      Good 150555
## 4      Fair  66726
## 5      Poor  27951
## 6      <NA>   1985
```

Here tabulating general_health data without “NA” values is being done

```
brfss2013 %>%
  filter(genhlth != "NA") %>%
  group_by(genhlth) %>%
  summarise(count = n())
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.3
```

```
## # A tibble: 5 x 2
##     genhlth  count
##      <fctr>  <int>
## 1 Excellent  85482
## 2 Very good 159076
## 3      Good 150555
## 4      Fair  66726
## 5      Poor  27951
```

Here tabulating medcost data is being done

```
brfss2013 %>%
  group_by(medcost) %>%
  summarise(count = n())
```

```
## # A tibble: 3 x 2
##   medcost  count
##    <fctr>  <int>
## 1     Yes  60107
## 2      No 430447
## 3    <NA>   1221
```

Here tabulating medcost data without “NA” values is being done

```
brfss2013 %>%
  filter(medcost != "NA") %>%
  group_by(medcost) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   medcost  count
##    <fctr>  <int>
## 1     Yes  60107
## 2      No 430447
```

Here tabulating (general_health & medcost data without “NA” values) are being done

```
brfss2013 %>%
filter(genhlth != "NA", medcost != "NA") %>%
  group_by(medcost, genhlth) %>%
  summarise(count=n())
```

```
## # A tibble: 10 x 3
## # Groups:   medcost [?]
##    medcost   genhlth  count
##     <fctr>    <fctr>  <int>
##  1     Yes Excellent   5166
##  2     Yes Very good  12449
##  3     Yes      Good  20295
##  4     Yes      Fair  14347
##  5     Yes      Poor   7493
##  6      No Excellent  80187
##  7      No Very good 146339
##  8      No      Good 129909
##  9      No      Fair  52140
## 10      No      Poor  20292
```

Here tabulating (general_health without “NA” values, medcost data without “NA” values & general_health is equal to poor) are being done

```
brfss2013 %>%
filter(genhlth != "NA", medcost != "NA", genhlth == "Poor") %>%
  group_by(medcost, genhlth) %>%
  summarise(count=n())
```

```
## # A tibble: 2 x 3
## # Groups:   medcost [?]
##   medcost genhlth count
##    <fctr>  <fctr> <int>
## 1     Yes    Poor  7493
## 2      No    Poor 20292
```

Conclusion

it can be concluded that no. of person who did not see the doctor due to cost having the poor health but this is not the only reason of poor general health, no. of people who see the doctor also having the poor health.

Research quesion 2:

Solution start from below

Here tabulating sleptim1 data is being done

```
brfss2013 %>%
  group_by(sleptim1) %>%
  summarise(count = n())
```

```
## # A tibble: 28 x 2
##    sleptim1  count
##       <int>  <int>
##  1        0      1
##  2        1    228
##  3        2   1076
##  4        3   3496
##  5        4  14261
##  6        5  33436
##  7        6 106197
##  8        7 142469
##  9        8 141102
## 10        9  23800
## # ... with 18 more rows
```

Here tabulating sleptim1 data without “NA” values is being done

```
brfss2013 %>%
  filter(sleptim1 != "NA") %>%
  group_by(sleptim1) %>%
  summarise(count = n())
```

```
## # A tibble: 27 x 2
##    sleptim1  count
##       <int>  <int>
##  1        0      1
##  2        1    228
##  3        2   1076
##  4        3   3496
##  5        4  14261
##  6        5  33436
##  7        6 106197
##  8        7 142469
##  9        8 141102
## 10        9  23800
## # ... with 17 more rows
```

Here tabulating qlhlth2 data is being done

```
brfss2013 %>%
  group_by(qlhlth2) %>%
  summarise(count = n())
```

```
## # A tibble: 25 x 2
##    qlhlth2 count
##      <int> <int>
##  1       0    97
##  2       1     6
##  3       2    20
##  4       3    11
##  5       4     6
##  6       5    25
##  7       6     1
##  8       7     6
##  9       8     5
## 10      10    16
## # ... with 15 more rows
```

Here tabulating qlhlth2 data without “NA” values is being done

```
brfss2013 %>%
  filter(qlhlth2 != "NA") %>%
  group_by(qlhlth2) %>%
  summarise(count = n())
```

```
## # A tibble: 24 x 2
##    qlhlth2 count
##      <int> <int>
##  1       0    97
##  2       1     6
##  3       2    20
##  4       3    11
##  5       4     6
##  6       5    25
##  7       6     1
##  8       7     6
##  9       8     5
## 10      10    16
## # ... with 14 more rows
```

Here tabulating (sleptim1 & qlhlth2 data without “NA” values) are being done

```
brfss2013 %>%
filter(sleptim1 != "NA", qlhlth2 != "NA") %>%
  group_by(sleptim1, qlhlth2) %>%
  summarise(count=n())
```

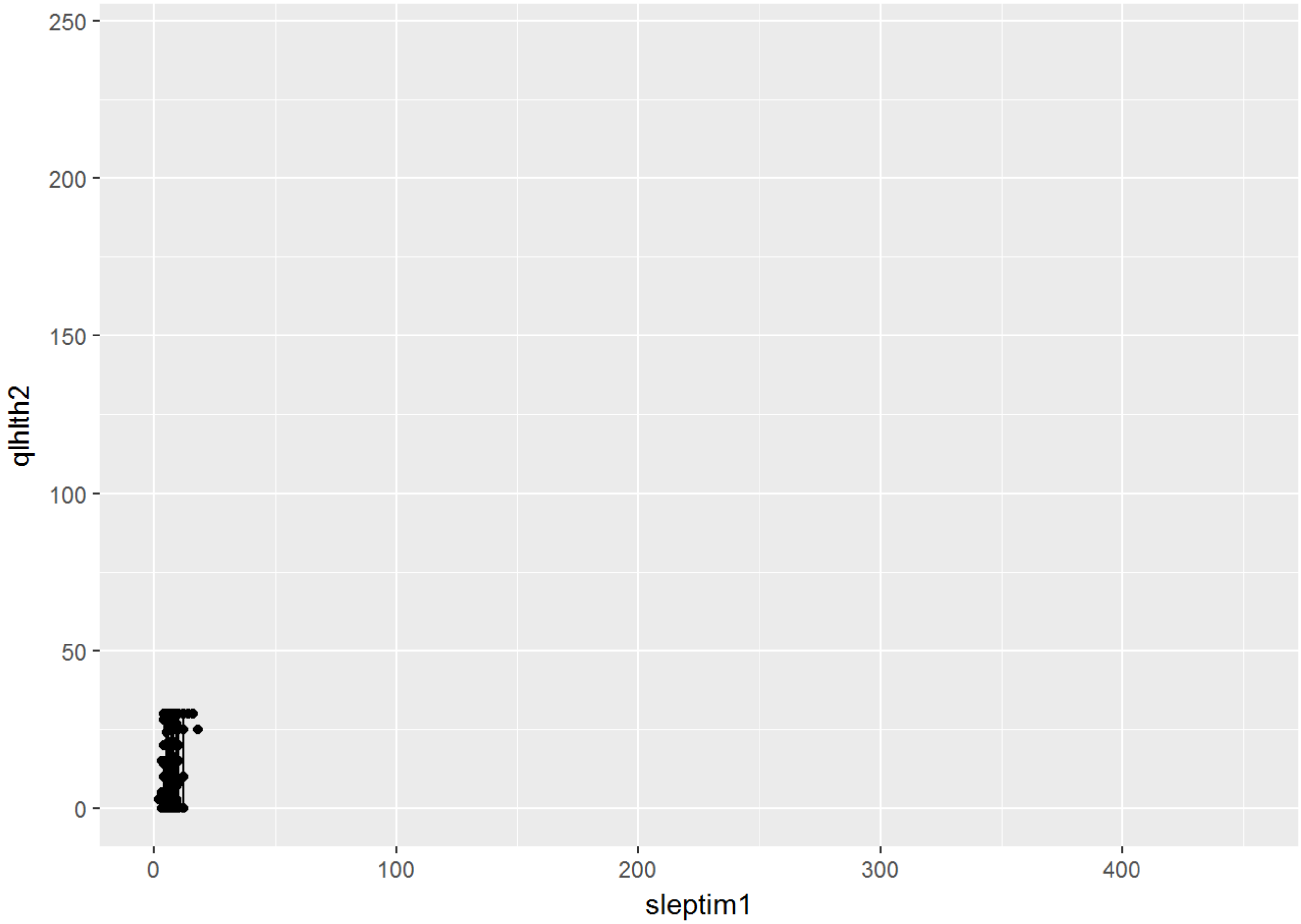
```
## # A tibble: 100 x 3
## # Groups:   sleptim1 [?]
##    sleptim1 qlhlth2 count
##       <int>   <int> <int>
##  1        2       3     1
##  2        3       0     3
##  3        3       5     1
##  4        3      15     1
##  5        4       0     8
##  6        4       1     2
##  7        4       3     2
##  8        4       4     1
##  9        4      10     1
## 10        4      14     1
## # ... with 90 more rows
```

I have included the visualization of sleeptime & healthyDays below.

```
ggplot(data = brfss2013, aes(x = sleptim1, y = qlhlth2)) +
  geom_line() +
  geom_point()
```

```
## Warning: Removed 7911 rows containing missing values (geom_path).
```

```
## Warning: Removed 491323 rows containing missing values (geom_point).
```



Here numeric operations applied on healthyDays

```
brfss2013 %>%
  summarise(qualtmean = mean(qlhlth2), qualtmedian = median(qlhlth2), qualtsd = sd(qlhlth2),
            qualtmin = min(qlhlth2), qualtmax = max(qlhlth2))
```

```
##   qualtmean qualtmedian qualtsd qualtmin qualtmax
## 1        NA          NA     NaN       NA       NA
```

Here numeric operations applied on healthyDays without “na” values

```
brfss2013 %>%
  filter(!(is.na(qlhlth2))) %>%
  summarise(qualtmean = mean(qlhlth2), qualtmedian = median(qlhlth2), qualtsd = sd(qlhlth2),
            qualtmin = min(qlhlth2), qualtmax = max(qlhlth2))
```

```
##   qualtmean qualtmedian  qualtsd qualtmin qualtmax
## 1  15.89032          15 15.93855        0      243
```

Below Computation:- if qlhlth2 is equal to or greater than “15” then “healthyDays” will be received otherwise Non healthy Dayswill be received

```
brfss2013 <- brfss2013 %>%
    filter(!(is.na(qlhlth2))) %>%
  mutate(healthyDays = ifelse(qlhlth2 >= 15, "healthyDays", "Not healthy Days"))
brfss2013 %>%
  group_by(healthyDays) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##        healthyDays count
##              <chr> <int>
## 1      healthyDays   267
## 2 Not healthy Days   198
```

Here numeric operations applied on sleptim1 variable without “na” values

```
brfss2013 %>%
  filter(!(is.na(sleptim1))) %>%
  summarise(sleptmean = mean(sleptim1), sleptmedian = median(sleptim1), sleptsd = sd(sleptim1),
            sleptmin = min(sleptim1), sleptmax = max(sleptim1))
```

```
##   sleptmean sleptmedian  sleptsd sleptmin sleptmax
## 1  7.073009           7 1.725375        2       18
```

if sleptim1 is equal to or greater than “15” without NA values then EnoughSlept will be received otherwise Not EnoughSleptime will be received

```
brfss2013 <- brfss2013 %>%
    filter(!(is.na(sleptim1))) %>%
  mutate(EnoughSlept = ifelse(sleptim1 >= 6, "EnoughSleptime", "Not EnoughSleptime"))
brfss2013 %>%
  group_by(EnoughSlept) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##          EnoughSlept count
##                <chr> <int>
## 1     EnoughSleptime   384
## 2 Not EnoughSleptime    68
```

if sleptim1 & qlhlth2 variables without “NA”, healthyDays is not equals to “Not healthy Days” then summary will be received grouped by healthyDays & EnoughSlept

```
brfss2013 %>%
  filter(!(is.na(sleptim1)), !(is.na(qlhlth2)), healthyDays != "Not healthy Days") %>%
  group_by(healthyDays,EnoughSlept) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 3
## # Groups:   healthyDays [?]
##   healthyDays        EnoughSlept count
##         <chr>              <chr> <int>
## 1 healthyDays     EnoughSleptime   242
## 2 healthyDays Not EnoughSleptime    20
```

Conclusion

it can be concluded that if person sleeps equal to or more than 6 hours then he will have healthy days or be full of energy. on the other hand a few amount of people also have healthyDays who have less than 6 hours sleep.

*__Research quesion 3:__

Solution start from below

Here tabulating exerany2 data is being done

```
brfss2013 %>%
  group_by(exerany2) %>%
  summarise(count = n())
```

```
## # A tibble: 3 x 2
##   exerany2 count
##     <fctr> <int>
## 1      Yes   272
## 2       No   177
## 3     <NA>     3
```

Here tabulating exerany2 data without “NA” values is being done

```
brfss2013 %>%
  filter(exerany2 != "NA") %>%
  group_by(exerany2) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   exerany2 count
##     <fctr> <int>
## 1      Yes   272
## 2       No   177
```

Here tabulating painact2 data is being done

```
brfss2013 %>%
  group_by(painact2) %>%
  summarise(count = n())
```

```
## # A tibble: 17 x 2
##    painact2 count
##       <int> <int>
##  1        0   295
##  2        1     9
##  3        2    16
##  4        3     9
##  5        4     5
##  6        5     7
##  7        6     3
##  8        7     4
##  9       10    10
## 10       12     3
## 11       15    10
## 12       20     8
## 13       21     2
## 14       25     5
## 15       30    56
## 16      206     1
## 17       NA     9
```

Here tabulating painact2 data without “NA” values is being done

```
brfss2013 %>%
  filter(painact2 != "NA") %>%
  group_by(painact2) %>%
  summarise(count = n())
```

```
## # A tibble: 16 x 2
##    painact2 count
##       <int> <int>
##  1        0   295
##  2        1     9
##  3        2    16
##  4        3     9
##  5        4     5
##  6        5     7
##  7        6     3
##  8        7     4
##  9       10    10
## 10       12     3
## 11       15    10
## 12       20     8
## 13       21     2
## 14       25     5
## 15       30    56
## 16      206     1
```

Here tabulating exerany2 & painact2 data without “NA” values are being done

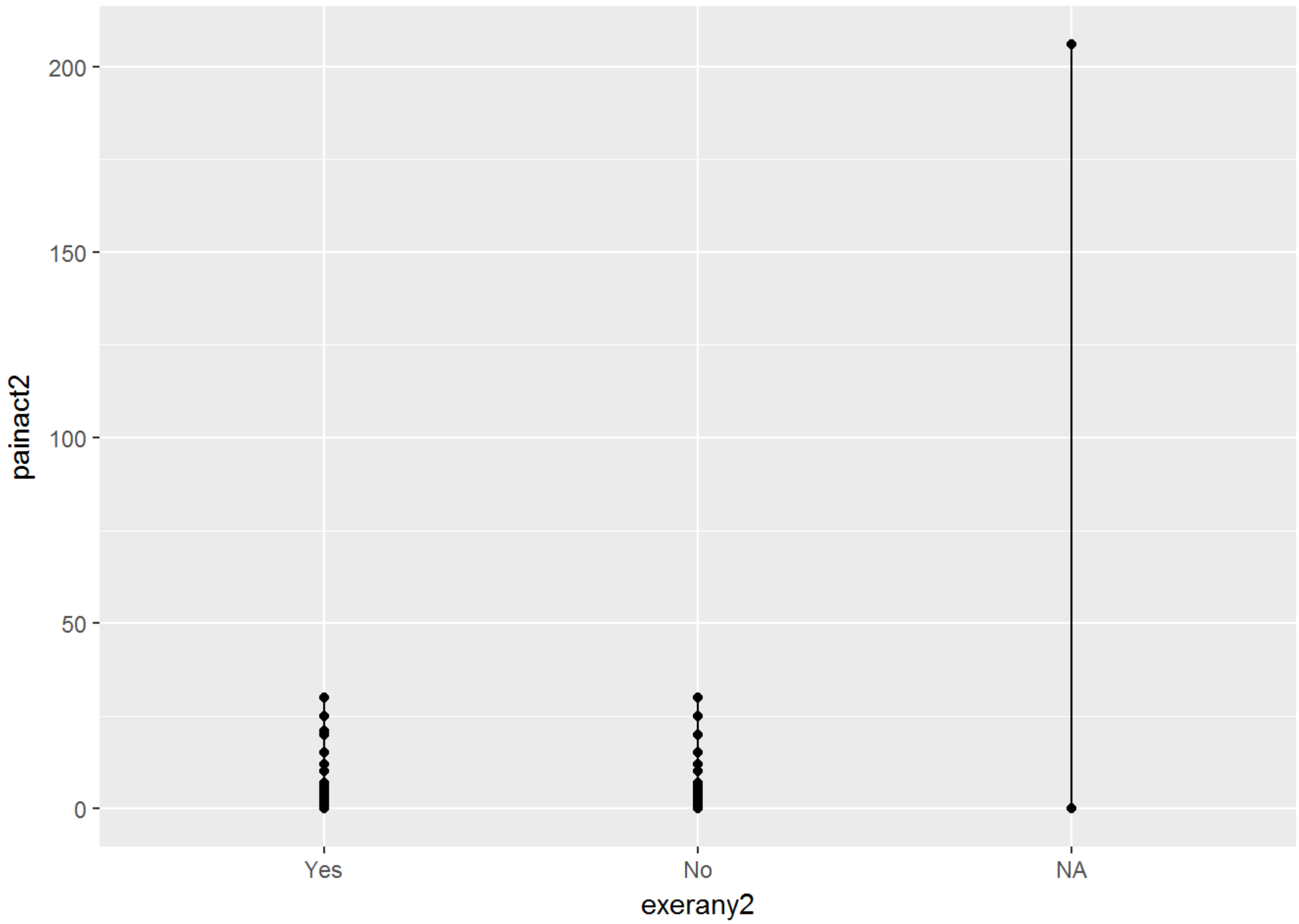
```
brfss2013 %>%
filter(exerany2 != "NA", painact2 != "NA") %>%
  group_by(exerany2, painact2) %>%
  summarise(count=n())
```

```
## # A tibble: 29 x 3
## # Groups:   exerany2 [?]
##    exerany2 painact2 count
##      <fctr>    <int> <int>
##  1      Yes        0   195
##  2      Yes        1     4
##  3      Yes        2     9
##  4      Yes        3     7
##  5      Yes        4     4
##  6      Yes        5     4
##  7      Yes        6     1
##  8      Yes        7     2
##  9      Yes       10     5
## 10      Yes       12     1
## # ... with 19 more rows
```

I have included the visualization of exerany2 & painact2 variable below. X axis contains exerany2 & Y axis contains painact2 variable

```
ggplot(data = brfss2013, aes(x = exerany2, y = painact2)) +
  geom_line() +
  geom_point()
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```



Here numeric operations applied on painact2 variable .

```
brfss2013 %>%
  summarise(painmean = mean(painact2), painmedian = median(painact2), painsd = sd(painact2),
            painmin = min(painact2), painmax = max(painact2))
```

```
##   painmean painmedian painsd painmin painmax
## 1       NA         NA    NaN      NA      NA
```

Here numeric operations applied on painact2 variable without “NA” values .

```
brfss2013 %>%
  filter(!(is.na(painact2))) %>%
  summarise(painmean = mean(painact2), painmedian = median(painact2), painsd = sd(painact2),
            painmin = min(painact2), painmax = max(painact2))
```

```
##   painmean painmedian   painsd painmin painmax
## 1 6.022573          0 14.13972       0     206
```

if painact2 is equal to or greater than “5” then HardActivities will be receivied otherwise Non HardActivities wil be received.

```
brfss2013 <- brfss2013 %>%
    filter(!(is.na(painact2))) %>%
  mutate(HardActivities = ifelse(painact2 >= 5, "HardActivities", "Not HardActivities"))
brfss2013 %>%
  group_by(HardActivities) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##       HardActivities count
##                <chr> <int>
## 1     HardActivities   109
## 2 Not HardActivities   334
```

if exerany2 is equal to “No” then Notexercized will be received otherwise Not exercized will be received on below tabulation

```
brfss2013 <- brfss2013 %>%
    filter(!(is.na(exerany2))) %>%
  mutate(Notexercized = ifelse(exerany2 == "No", "Notexercized", "exercized"))
brfss2013 %>%
  group_by(Notexercized) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   Notexercized count
##          <chr> <int>
## 1    exercized   268
## 2 Notexercized   172
```

if exerany2 & painact2 variables without “NA” values, Notexercized is not equal to “exercized” then tabulate both of them.

```
brfss2013 %>%
  filter(!(is.na(exerany2)), !(is.na(painact2)), Notexercized != "exercized") %>%
  group_by(Notexercized, HardActivities) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 3
## # Groups:   Notexercized [?]
##   Notexercized     HardActivities count
##          <chr>              <chr> <int>
## 1 Notexercized     HardActivities    59
## 2 Notexercized Not HardActivities   113
```

conclusion

It can be seen that count of the no. of Notexercized person having not Hardactivities is greated than no. of Notexercized person having Hardactivities, Therefore it can be concluded that some persons who Notexercized have HardActivities but more amount of persons who Notexercized do not have HardActivities.

End of solution