Author:Sanjoy Basu

Exploring the BRFSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(reshape2)
library(gridBase)
library(gridExtra)
```

Load data

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

Part 1: Data

The research is using 2013 data collected by BRFSS. The Behavioral Risk Factor Surveillance System (BRFSS) with technical and methodological assistance from CDC, state health departments conduct health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors. The survey is conducted using Random Digit Dialing (RDD) techniques on both landlines and cell phones. So the data is from randomly sampled subjects from the population of United States. This data is generalizable of US population and can be used only for observational study.

Part 2: Research questions

Research quesion 1: Is there a relation between reported number of days of poor health condition (physical and mental) and lack of medical coverage?

Background and motivation Health insurance mandate also known as Patient Protection and Affordable Care Act (ACA) has been the eye of political storm in US. Opposition to a government role in mandatory health insurance argues that it is economically unsustainable for US to be able to insure all its citizen. On the other hand according to an article in Forbes by Bruce Japsen dated Sep12 2012 The Integrated Benefits Institute, which represents major U.S. employers and business coalitions, says poor health costs the U.S. economy $576 billion a year, according to new research. Of that amount, 39 percent, or $227 billion is from “lost productivity” from employee absenteeism due to illness. With the

above information and BRFSS data I would like to establish if there is a relation between number of days lost due to poor health and lack of medical coverage. This is done by estimating total number of days residents of a state or territory report poor health condition and total number of residents in the state with no or very limited health coverage.

Research quesion 2: Is there a relation between good health and consumption of food such as fruits and dark-green, orange vegetables ?

Background and motivation Food and Drug Administration nutrition expert (FDA's) Barbara Schneeman recommends: "Eat at least 4.5 cups of fruits and vegetables a day, including a variety of dark-green, red, and orange vegetables, beans, and peas." I try to answer the question by estimating number of subjects who consider that they have excellent health condition and their consumption of fruits and vegetables vs those who says that they have poor heath condition and their consumption of fruits and vegetables.

Research quesion 3: Is there a relation between intensity of depression and number of hours of sleep?

Background and motivation According to National Sleep Foundation insomnia is very common among depressed patients. Feeling sad and depressed is part of human experience but intensity of depression is the focus of my reasearch. Using BRFSS data for 2013 I would like to establish if there is a correlation between intensity of depression subject report and the number of hours they sleep.

Part 3: Exploratory data analysis

Research quesion 1:

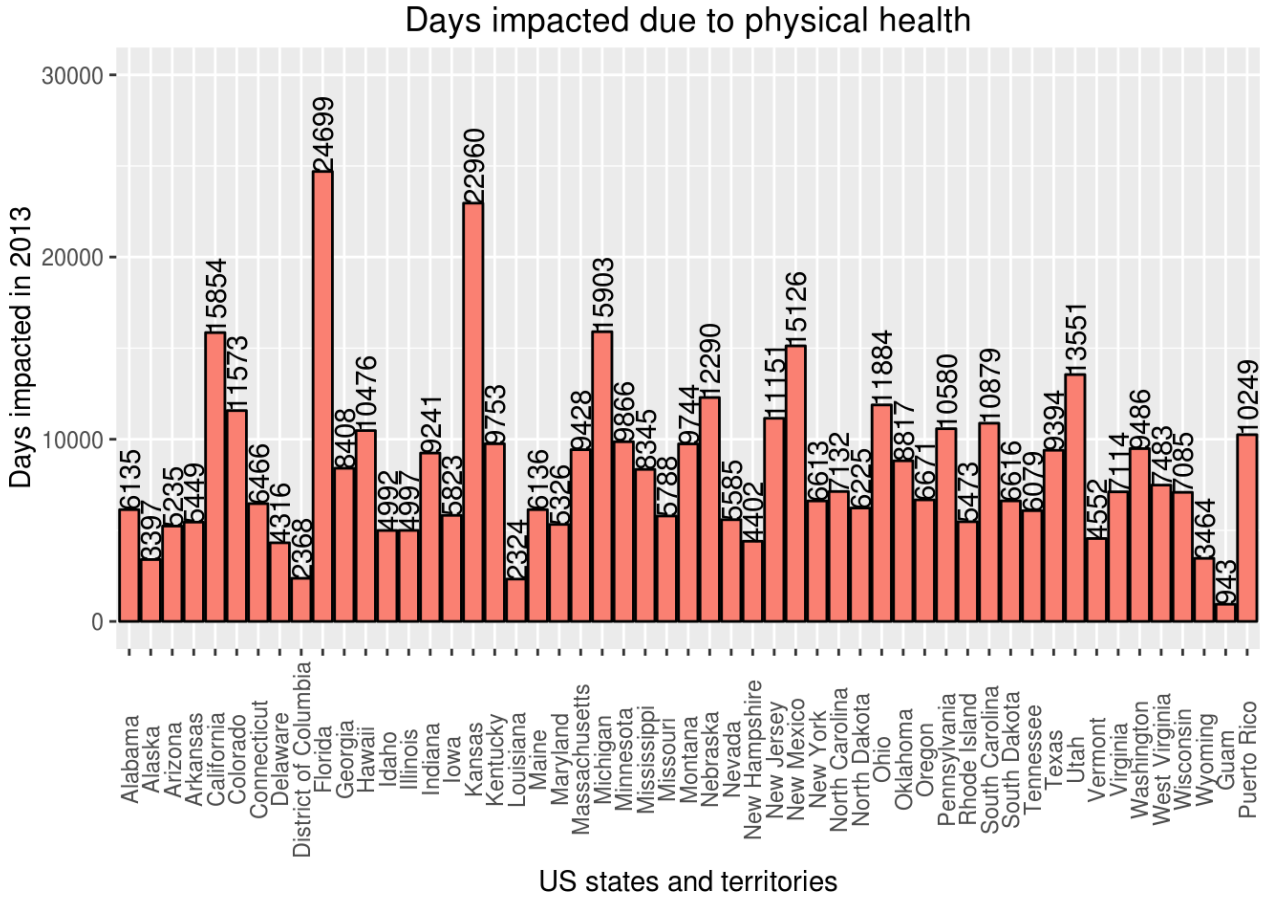
```
phys_ment_statedistribution <- brfss2013 %>% filter(!is.na(cstate), !is.na(phys
hlth), !is.na(menthlth), cstate=="Yes") %>% group_by(X_state) %>% select (X_sta
te, physhlth, menthlth) %>% summarise(phys=sum(physhlth), ment=sum(menthlth))
head(phys_ment_statedistribution, n=10)
```

```
## Source: local data frame [10 x 3]
##
##                 X_state  phys  ment
##                  <fctr> <int> <int>
## 1               Alabama  6135  6580
## 2                Alaska  3397  3283
## 3               Arizona  5235  4856
## 4              Arkansas  5449  5394
## 5            California 15854 17854
## 6              Colorado 11573 12325
## 7           Connecticut  6466  7322
## 8              Delaware  4316  4608
## 9  District of Columbia  2368  2873
## 10              Florida 24699 25331
```

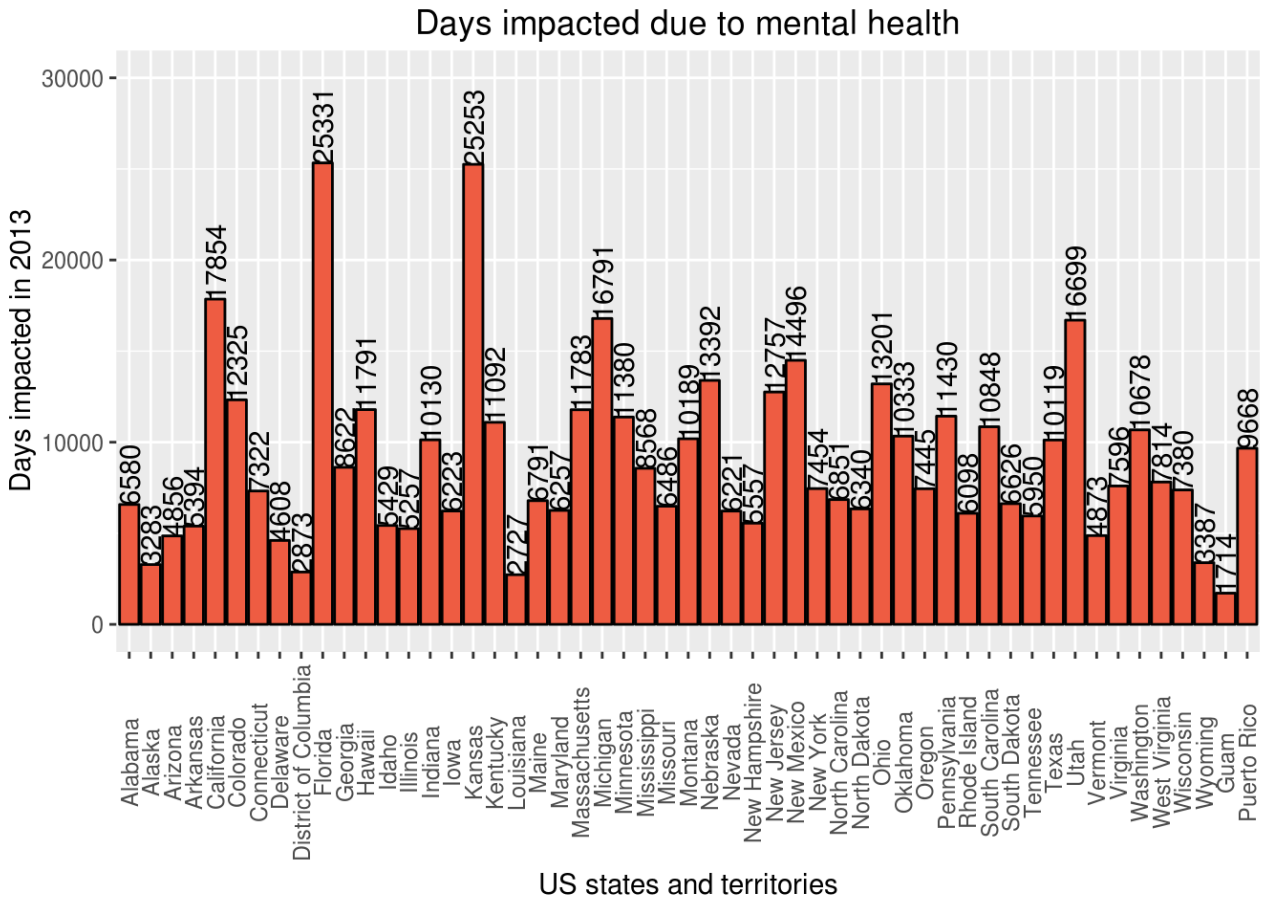
```
state_cover <- brfss2013 %>% filter(!is.na(cstate), cstate=="Yes", !is.na(nocov
121))  %>% group_by(X_state)  %>%  summarise(cnt=sum(nocov121=="Yes"))
head(state_cover, n=10)
```

```
## Source: local data frame [10 x 2]
##
##                 X_state   cnt
##                  <fctr> <int>
## 1               Alabama    91
## 2                Alaska    80
## 3               Arizona   103
## 4            California   322
## 5           Connecticut   115
## 6              Delaware    80
## 7  District of Columbia    65
## 8               Florida   443
## 9               Georgia   165
## 10                Idaho   112
```

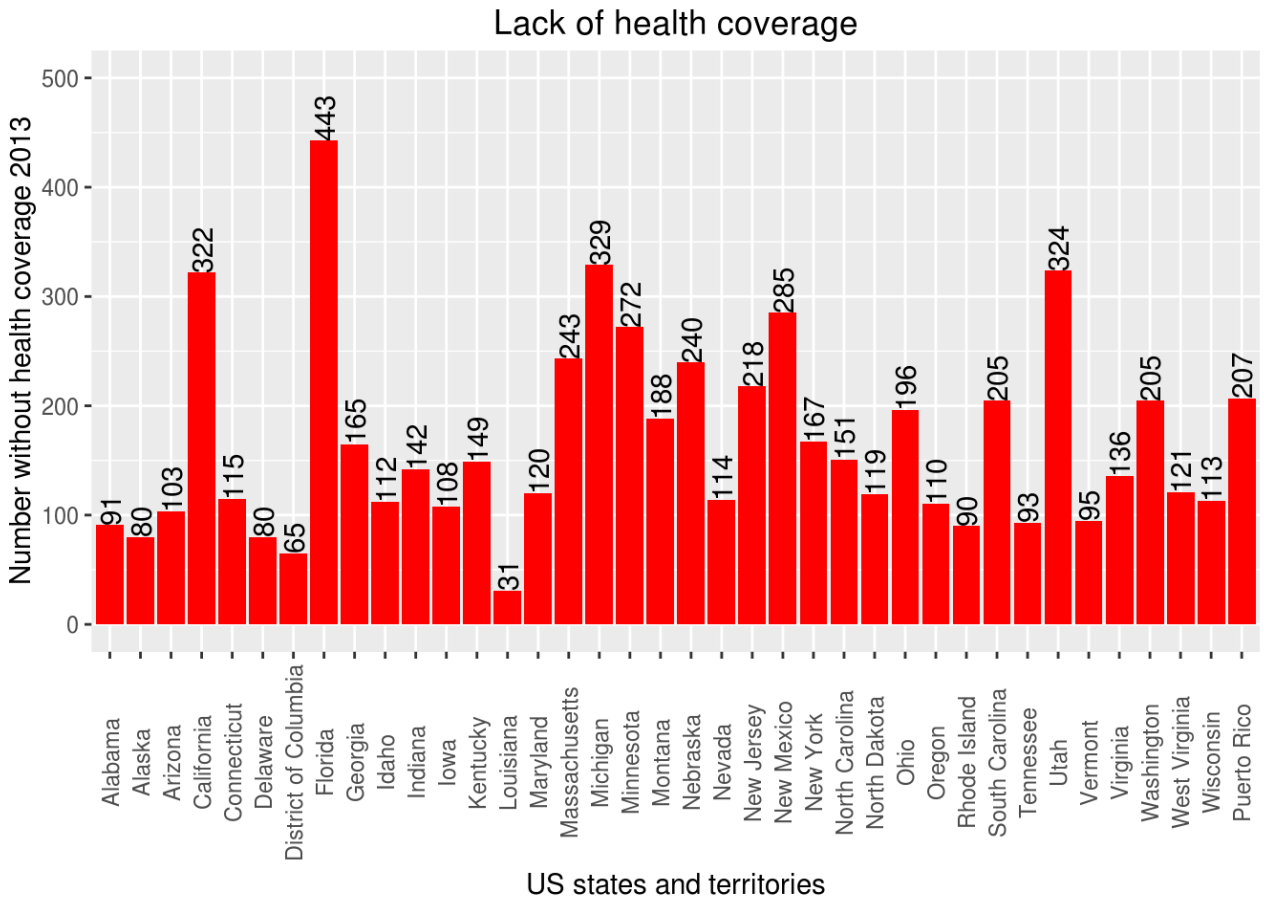
```
phys_plot <- ggplot(data=phys_ment_statedistribution, aes(x=X_state, y=phys))+g
eom_bar(color="black", stat = "identity", fill="salmon") + scale_y_continuous(l
imits=c(0, 30000)) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
 + xlab("US states and territories") + ylab("Days impacted in 2013") + geom_tex
t(aes(label=phys, angle=90, hjust=0.019)) +  ggtitle("Days impacted due to phys
ical health")
mntl_plot <- ggplot(data=phys_ment_statedistribution, aes(x=X_state, y=ment))+g
eom_bar(color="black", stat = "identity", fill="tomato2") + scale_y_continuous(
limits=c(0, 30000)) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5)
) + xlab("US states and territories") + ylab("Days impacted in 2013") + geom_te
xt(aes(label=ment, angle=90, hjust=0.019)) +  ggtitle("Days impacted due to men
tal health")
cvrg_plot <- ggplot(data=state_cover, aes(x=X_state, y=cnt))+ geom_bar(stat="id
entity", fill="red") + scale_y_continuous(limits=c(0, 500)) + theme(axis.text.x
 = element_text(angle = 90, vjust = 0.5)) + xlab("US states and territories") +
 ylab("Number without health coverage 2013") + geom_text(aes(label=cnt, angle=9
0, hjust=0.019)) +  ggtitle("Lack of health coverage")
#grid.arrange(phys_plot, mntl_plot, nrow=2)
phys_plot
```



```
mntl_plot
```




```
cvrg_plot
```

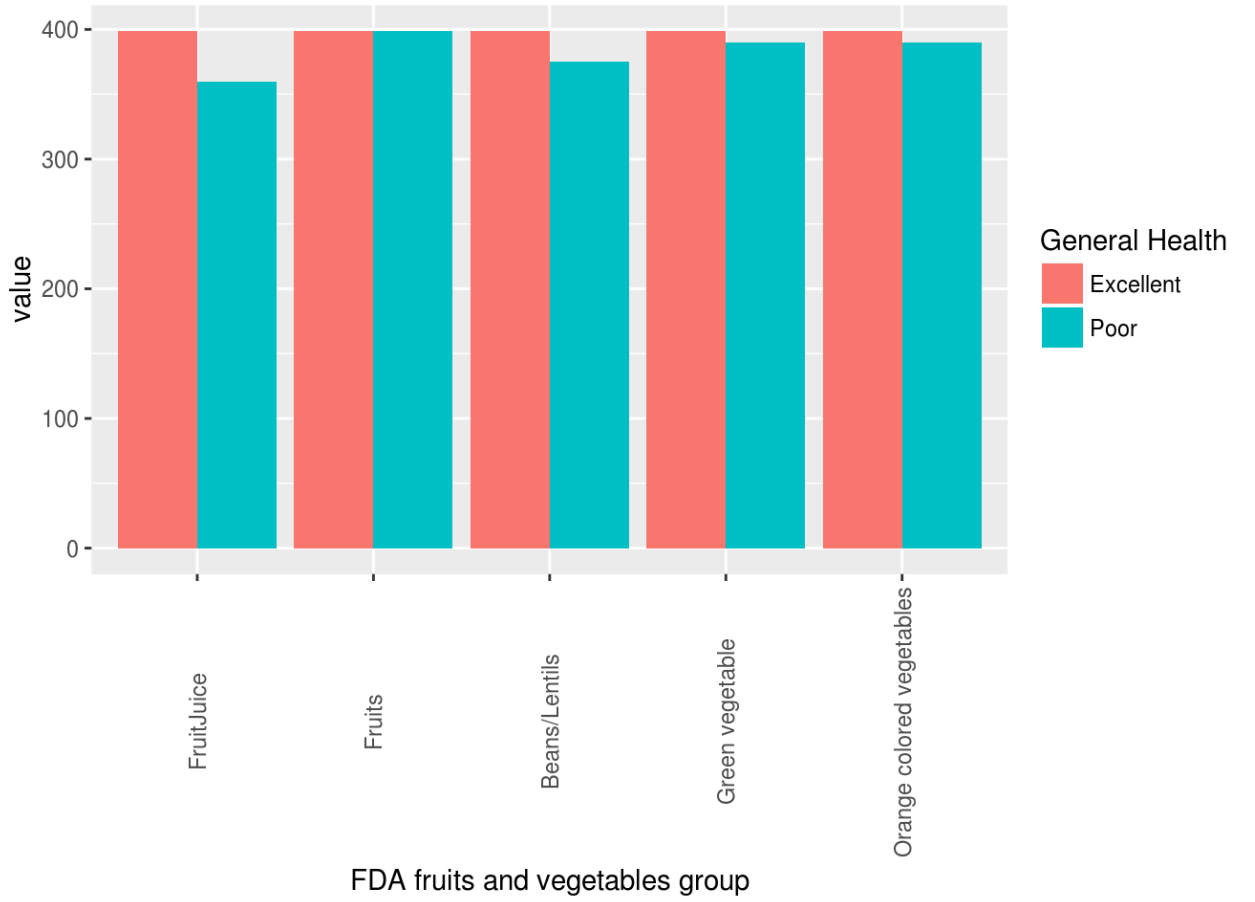


It can be clearly seen that number of residents of a state who has no or very limited access to health coverage is proportional to the number of days impacted due to poor physical and mental health. For example resident of Florida has highest number of days of reported poor health and Florida has also highest number of residents with no or very limited health coverage followed by Michigan (health coverage data not available for Kansas). This suggest that lack of health coverage may have impact on productivity of a state or a territory and subsequently on its economy.

Note: I understand absolute number of days and number of residents with no or limited access to healthcare coverage is not suitable for comparison as each state has different polpulation . Percentage is more suitable for such comparision. Unfortunately BRFSS data does not provide population information of each state required for calculation of percentage. Percentage can be calculated by combining data from other sources such as US Census Bureau. Scope of the research is limited to BRFSS data of 2013.

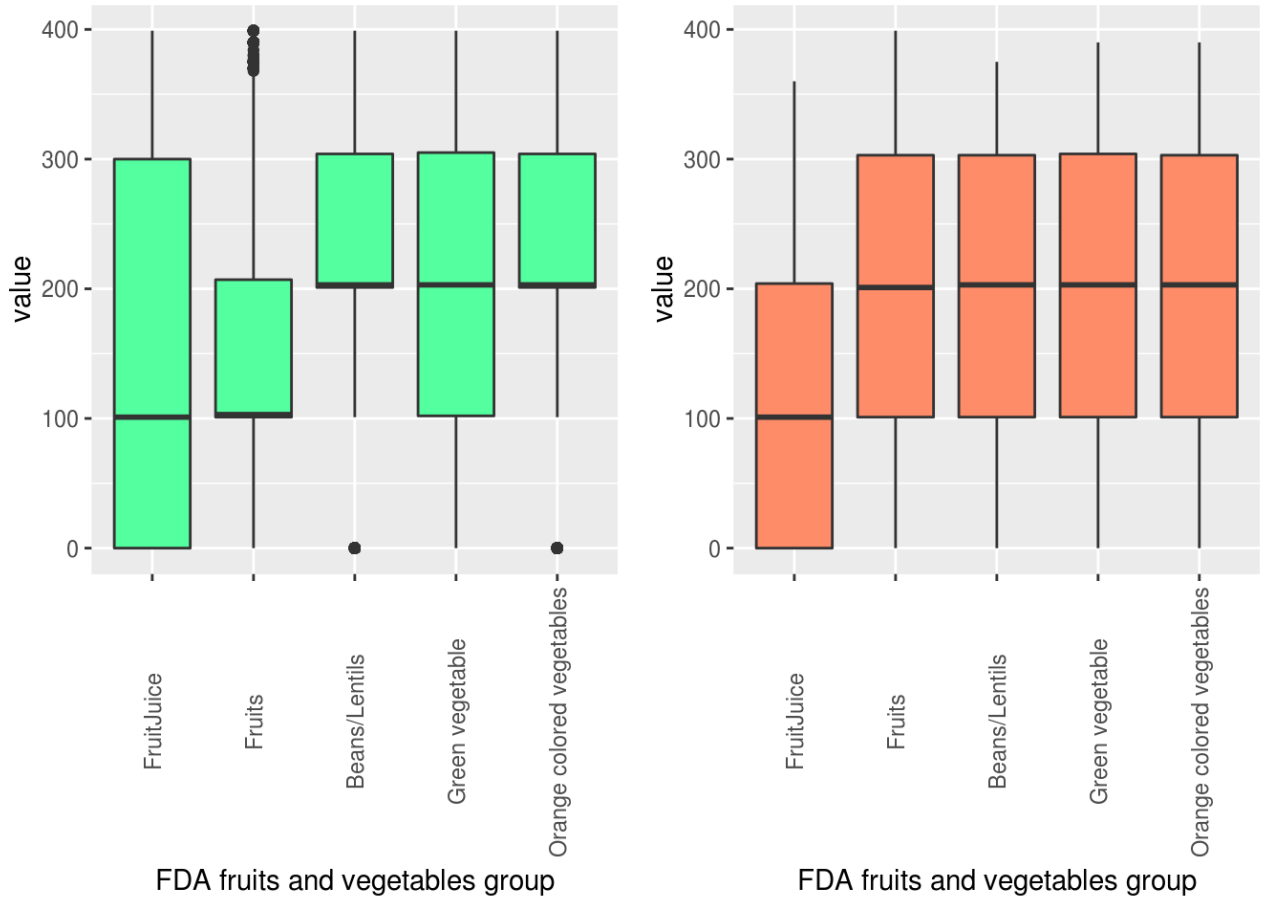
Research quesion 2:

```
genhlth_food <- brfss2013  %>% filter(!is.na(genhlth),  !is.na(fruitju1), !is.n
a(fruit1),!is.na(fvbeans),!is.na(fvgreen),!is.na(fvorang)) %>% select (genhlth,
 fruitju1, fruit1, fvbeans, fvgreen, fvorang)
genhlth_foodm<-melt(genhlth_food, id.vars = "genhlth")
genhlth_foodmpe <- genhlth_foodm %>%  filter(genhlth=="Excellent" | genhlth=="P
oor")
gf_bar <- ggplot(data=genhlth_foodmpe, aes(variable, value, fill=genhlth)) + ge
om_bar(stat = "identity", position="dodge") + scale_fill_discrete(name="General
 Health") + xlab("FDA fruits and vegetables group")
gf_bar<-gf_bar + scale_x_discrete(labels=c("FruitJuice", "Fruits", "Beans/Lenti
ls", "Green vegetable", "Orange colored vegetables")) + theme(axis.text.x = ele
ment_text(angle = 90, vjust = 0.5))
gf_bar
```



The bar graph suggests that subjects with excellent health usually consume more fruits and dark-green, orange vegetables than those with poor health. To get more insight of their consumption habit we use 2 seperate boxplots : green for subjects with excellent health and orange for subjects with poor health.

```
genhlth_foodmp <- genhlth_foodm %>% filter(genhlth=="Poor")
genhlth_foodme <- genhlth_foodm %>% filter(genhlth=="Excellent")
s1 <- ggplot(data=genhlth_foodme, aes (x=variable, y=value)) + geom_boxplot(fil
l="seagreen1") + xlab("FDA fruits and vegetables group") + scale_x_discrete(lab
els=c("FruitJuice", "Fruits", "Beans/Lentils", "Green vegetable", "Orange color
ed vegetables"))
s2 <- ggplot(data=genhlth_foodmp, aes (x=variable, y=value)) + geom_boxplot(fil
l="salmon1") + xlab("FDA fruits and vegetables group") + scale_x_discrete(label
s=c("FruitJuice", "Fruits", "Beans/Lentils", "Green vegetable", "Orange colored
 vegetables"))
s1 <- s1 + theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
s2 <- s2 + theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
grid.arrange(s1, s2, ncol=2)
```



Fruit Juice: Box plot indicates the distribution of consumption is right skewed for subjects with excellent health which means while each group share same median on average subjects with excellent health consume more fruit juice. Also 75th percentile of subjects with excellent health consume more fruit juice than subjects with poor health.

Fruits: Mean and median of consumption raw fruit is lower for subjects with excellent health also 75th percentile of subjects with poor health consume more fruits. This apperently contradicts FDA and health

care professional's reccomendation. However this does not suggest that more fruit consumption is causation for poor health as this is merely a correlation. Also it can be noted raw fruits are perishable and organic and are vulnerable to rottening and bacterial growth. Confounding variables such as quality (fresh vs rotten) and hygenic condition is beyond the scope of BRFSS data.

Beans and Lentils: While both group shares same median, on average healthy subjects consume more beans and lentils (distribution right skewed). Also 25th percentile of subjects with excellent health consume more beans and lentils than subjects with poor health.

Green vegetables: No visble difference in consumption between the group. This does not suggest that there is no relation between consumption of green vegetable and good health. Also it can be noted vegetables are usually consumed raw and/or cooked. Raw vegetables are perishable and organic and are vulnerable to rottening and bacterial growth. Confounding variables such as quality (fresh vs rotten), consumption method (cooked vs raw) and hygenic condition is beyond the scope of BRFSS data.

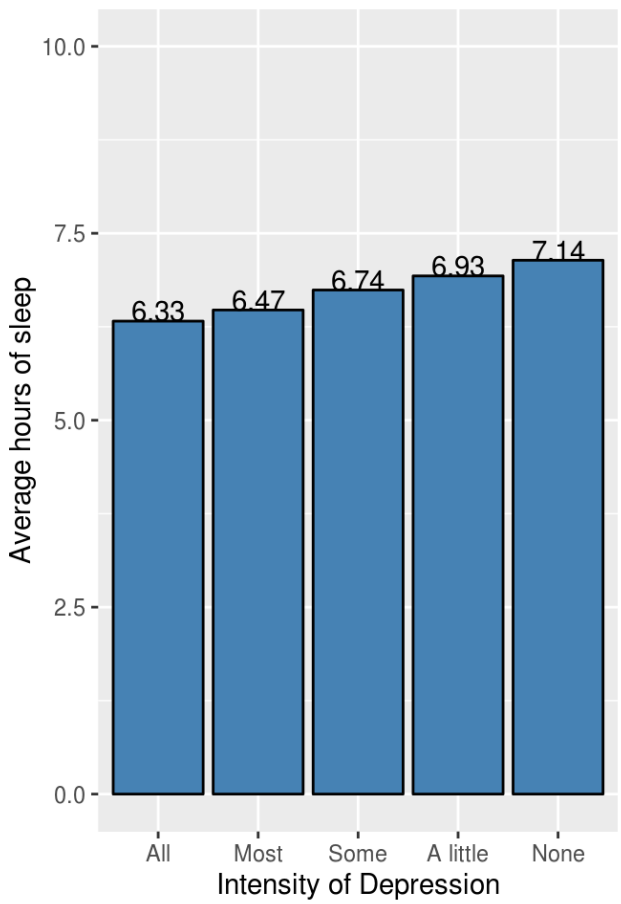
Orange colored vegetables: While both group shares same median, on average healthy subjects consume more orange colored vegetables (distribution right skewed). Also 25th percentile of subjects with excellent health consume more orange color vegetables than subjects with poor health.

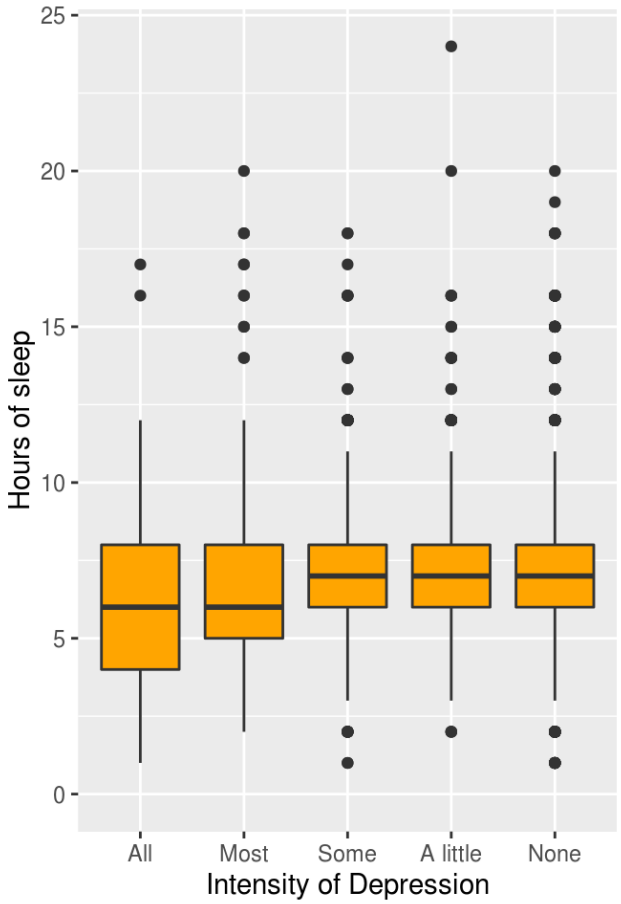
Above research shows a trend that subjects with excellent health consume more fruits and dark-green, orange vegetables however from observation it remains inconclusive if consumption of fruits and dark-green, orange vegetables is cause for good health. There appear to be confounding variables such as quality (fresh vs rotten), consumption method (cooked vs raw) and hygenic condition that may affect correlation. Also experiment with random assignment of subjects is preferable over observational study for accurate conclusion.

Research quesion 3:

```
slp_mean <- brfss2013 %>% filter(!is.na(sleptim1), !is.na(misdeprd))  %>% group
_by(misdeprd) %>% summarise(sleptim1_mean=mean(sleptim1)) %>% select (sleptim1_
mean, misdeprd)
slp <- brfss2013 %>% filter(!is.na(sleptim1), !is.na(misdeprd)) %>% select (sle
ptim1, misdeprd)
d1 <- ggplot(data=slp_mean, aes(x=misdeprd, y=sleptim1_mean)) + geom_bar(stat="
identity", color="black", fill="steelblue") + scale_y_continuous(limits = c(0,
10)) + geom_text(aes(label=round(sleptim1_mean, 2), , vjust=0.019)) + xlab("Int
ensity of Depression") + ylab("Average hours of sleep")
d2 <- ggplot(data=slp, aes(x=misdeprd, y=sleptim1)) + geom_boxplot(fill="orange
") + scale_y_continuous(limits = c(0, 24)) +xlab("Intensity of Depression") + y
lab("Hours of sleep")
grid.arrange(d1, d2, ncol=2)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```





The bar graph indicates that average hours of sleep is less for subjects with severe depression than subjects with mild or no depression. However the difference is too small. Box plot gives us better insight of sleepng habit of subjects with different degrees of depression. It is clear that median and mean hours of sleep is less for subjects with severe depression in comparision to subjects with mild or no depression. We also see that 25 percentile of subjects who suffer from severe depression has less hours of sleep than 25 percentile of those who has less severe depression. IQR for sleeping hours of subjects with mild or no depression is narrower than subjects with severe depression which suggest there is likely sleeping disorder among subjects with severe depression , this is also supported by length of whiskers of box plots representing subjects with different severity of depression.

Statistical evidence suggest that there is a corelation between intensity of depression and hours of sleep. This is however not sufficient to diagnose level of depression as correlation suggest mere generalization and cannot be used as causation. Experiment with random assignment of subjects is preferable over observational study for accurate conclusion.