Exploring the BRFSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
```

Load data

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `brfss2013`. Delete this note when before you submit your work.

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

Part 1: Data

The BRFSS's objective is to collect data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries and preventable infectious diseases that affect the adult population. The meticulous data collected by BRFSS is then used by statisticians for study and to provide valuable causations and results between certain variables included in the BRFSS data. However for these studies and results to be generated, the data should be generalizable or in other words should be collected in a way that the given sample of data represents a population.

Reasons For Generalizability:

- Questions Asked: There are 4 types of questions asked
  - Annual Standard core: A set of standardized questions that every state is supposed to include in its questionnaire.
  - Rotating Core Questions: These questions are asked by all the states on every other year basis.
  - Optional Modules: These modules contain sets of standardized questions on various topics that each state may select and include in its questionnaire.
  - State Added Questions: Depending on the state's specific health priorities, the state may choose to add certain questions to the questionnaire.

Thus as we can see, the four types of questions are predefined by the BRFFS, and the state can select their questions only from those

questions and cannot make up their own questions.Thus there is a good generalizability when it comes to the questionnaire.

- Sampling Design: There are two types of samples:
  - Landline Sample: Disproportionate Stratified Sampling(DSS) has been used for landline sample .DSS draws numbers from two strata, which are based on the density of listed and non-listed numbers. This DSS design finds a way of differentiating before sampling begins. Since landlines are connected to a household, the interviewers collect information on the number of adults living within the residence and then select randomly from all the eligible bachelors.

– Cellular Telephone Sample: The cellular telephone respondents are selected randomly with each having equal probability of selection.

Thus we can see from the above two sampling methods that the data acquired is purely random in nature and is representative of the
population as a whole.

- Geographic Stratification: The BRFSS samples landline telephone numbers on sub-state geographic regions. The data is collected as a representation of geographic strata. This takes time and money but , data collected in this way is representative of the regional strata and in the end of the population as a whole.
- Weighting: The data provided by BRFSS is weighted data, which attempts to remove bias in the data. The weighting protocols of BRFFS has ensured that data are representative of the population on a number of demographic characteristics including sex, age, race, education, marital status etc.

Causality:

- The questions asked are general
- The households are stratified and then using random sampling adults are chosen from the household and interviewed. Also students living in the University household are considered as single households, i.e. only one person lives in the household.
- The cellular telephone sample is randomly generated from a sampling frame of confirmed cellular area code and prefix combinations. After that cellular telephone respondents are randomly selected for interview.

Part 2: Research questions

- Research question 1:
  - Question: Is a person's sleep related to his or her general health, that is does a person who sleeps an average of 7 to 9 hours have a better health in general than a person who sleeps above or below the average ?
  - Reason for the question: Sleep is an essential part of our day to day life. A lot of doctors ask us to have a good night's sleep of about 6.5 to 9 hours for the proper functioning of our body.It is said that the above mentioned amount of sleep is necessary for good, mental and physical health.It is also stated by lot of wellness organizations that proper sleep is necessary for the body's immune system to be restored, the tissues to get repaired etc. All these benifits of proper sleep lead us to believe that if a person sleeps properly his or her's general health is better than the people who over sleep or under sleep. And it is this belief that I want to answer, i.e. does proper sleep actually affect your health and if so, do the people having the average amount of sleep i.e 7 to 9 hours have better health than others ?
  - Variables used:

    -> genhlth: Variable corrosponding to health of a person
    -> sleptim1: How many hours a person sleeps in 24 hours on an average

- Research question 2:
  - Question: How is a person's health related to his or her's income level and health plan? Also do people with higher income tend to have health plans which helps them with better health or is there no correlation between health plan and general health?Also depending on the analysis should a person irrespective of his or her's income buy a health plan ?

  - Reason for the question: There are a lot of health plans today in United States. But there is a downside to it, they are costly and there are times when people with a lesser salary cannot afford to buy a health plan. However if a person falls sick and he or she does not have a helath plan, the medical cost for treating him or her could be much higher than the actual money he would be paying for the health plan. In such a situation a person with a lesser salary would have been better off if he or she would have bought a health plan. Hence the reason for this analysis is to find the importance of Health plans for a person's health.
  - Variables used:

    -> genhlth: Variable corrosponding to health of a person
    -> income2: Specifies the income level of a person.
    -> hlthpln1: Specifies if a person has a health plan or not.

- Research question 3:
  - Question: How is a person's emotional support related to the satisfaction of his or her's life? Also is it true that females need more emotional support as compared to males in order to be satisfied with life?
  - Reason for the question: We have often heard that a person committed suicide because he or she was not happy with his or her life. Now a person may not be happy with his or her life if something bad has happened or due to some medical conditions.In such perilious times what a person would need is support from his or her's close ones. I want to explore in this data analysis if emotional support actually corrosponds with a better satisfaction towards life.Also in the wake of 21st centuary it is said that females are emotionally more vulnerable than man, however there is no substance to it. It would be good if we do some data analysis and to find this thing out and check whether emotional support correlates to a better satisfaction towards life irrespective of the gender.
  - Variables used:

    -> emtsuprt: Level of emotional satisfaction that a person receives
    -> lsatisfy: The amount of satisfaction that a person has towards life
    -> sex: The sex of a person

Part 3: Exploratory data analysis

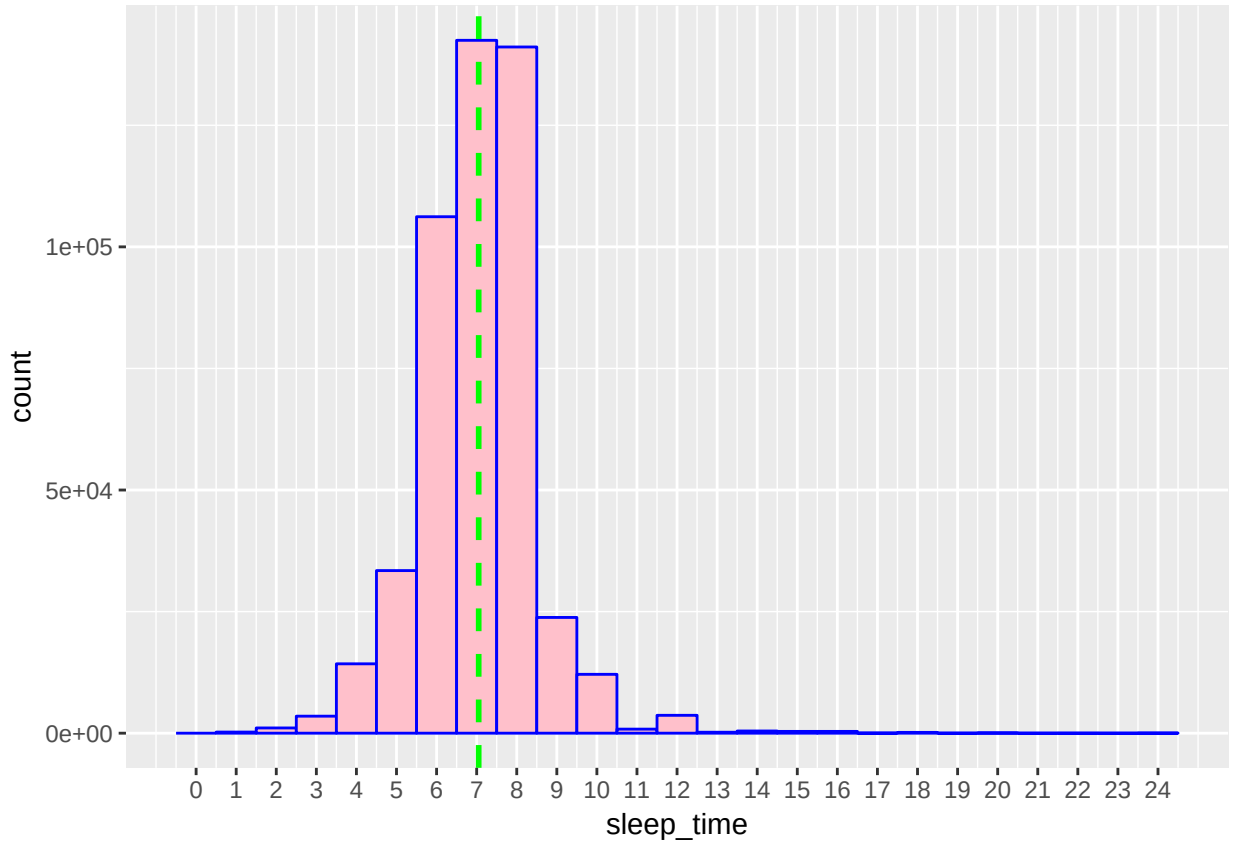
Research quesion 1:

- Cleaning the Data:
  - There are 24 hours in a day , but the sleeptim1 feature consisted some unnecessary values like 103 and 450, which are definitely due to the data entering error. As a result we need to remove such unnecessary values.

```
brfss2013$sleptim1[as.numeric(brfss2013$sleptim1) > 24] <- NA
sleep_time <- brfss2013$sleptim1
```

```
mean_hrs <- mean(as.numeric(sleep_time),na.rm=TRUE)
stdev_hrs <- sd(as.numeric(sleep_time),na.rm=TRUE)
p <- ggplot(data = brfss2013,aes(x=sleep_time))+geom_histogram(color="blue",fill="pink",binwidth = 1)
p <- p + geom_vline(aes(xintercept=mean_hrs),color="green", linetype="dashed",size=1)
```

```
p <- p+scale_x_continuous(breaks = round(seq(min(sleep_time,na.rm = TRUE), max(sleep_time,na.rm = TRUE)
p
```



- As you can see, the above graph is a histogram of average sleeping time of people in the data set. Its clearly visible that the above graph follows a normal distribution.
- The dashed green line in the above figure depicts the mean of the plot. As mentioned below the mean is 7.050986 and the standard deviation is 1.465987.
- Since the above graph resembles a normal distribution according to the 68-96-99.7 rule, almost 68% of the people lie withing the first standard deviation of the mean. This is to say that almost 68% of the people sleep in the range of 5.58 hours to 8.52 hours.
- As mentioned in the question doctors suggest people to sleep within 6.5 to 9 hours daily for good health. This range is almost equal to the range of sleep 68% of the people in BRFSS datset sleep.

```
mean_hrs
```

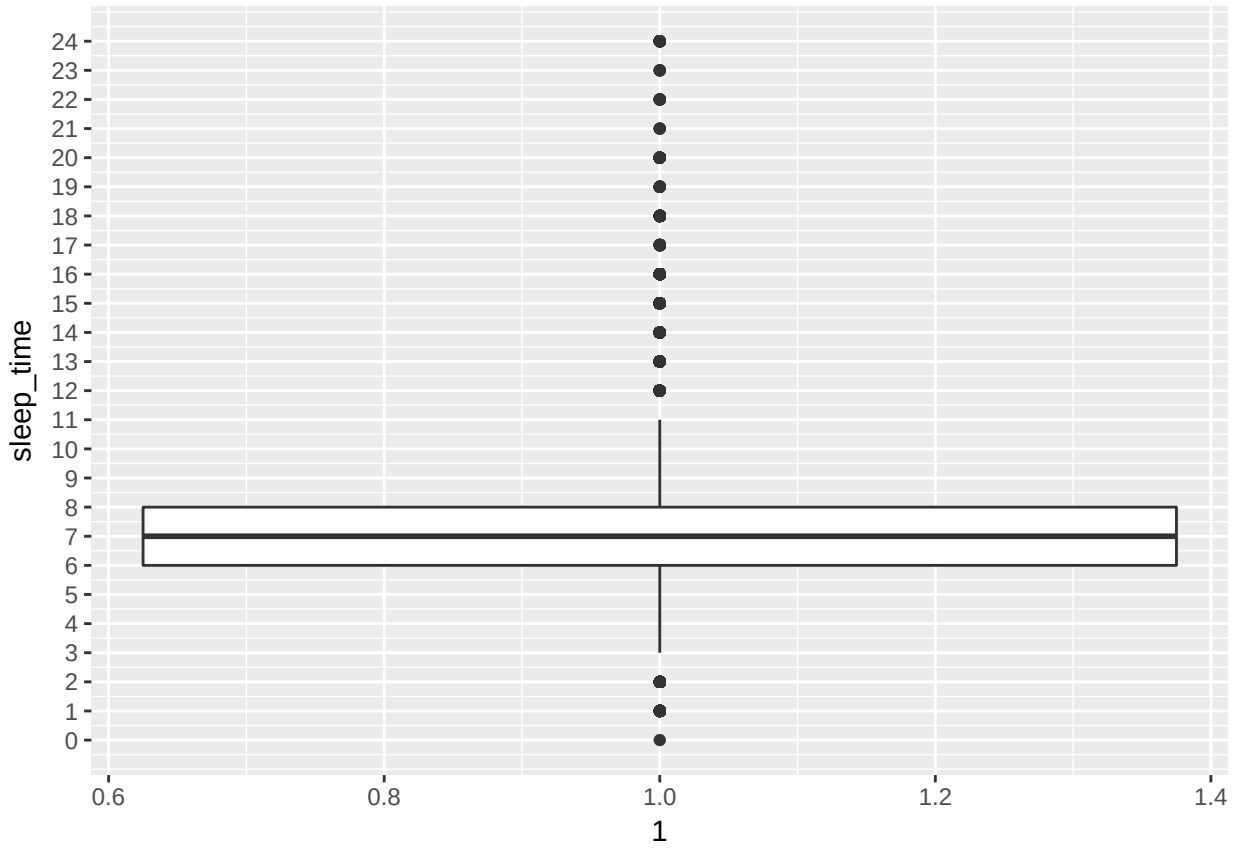
```
## [1] 7.050986
```

```
stdev_hrs
```

```
## [1] 1.465987
```

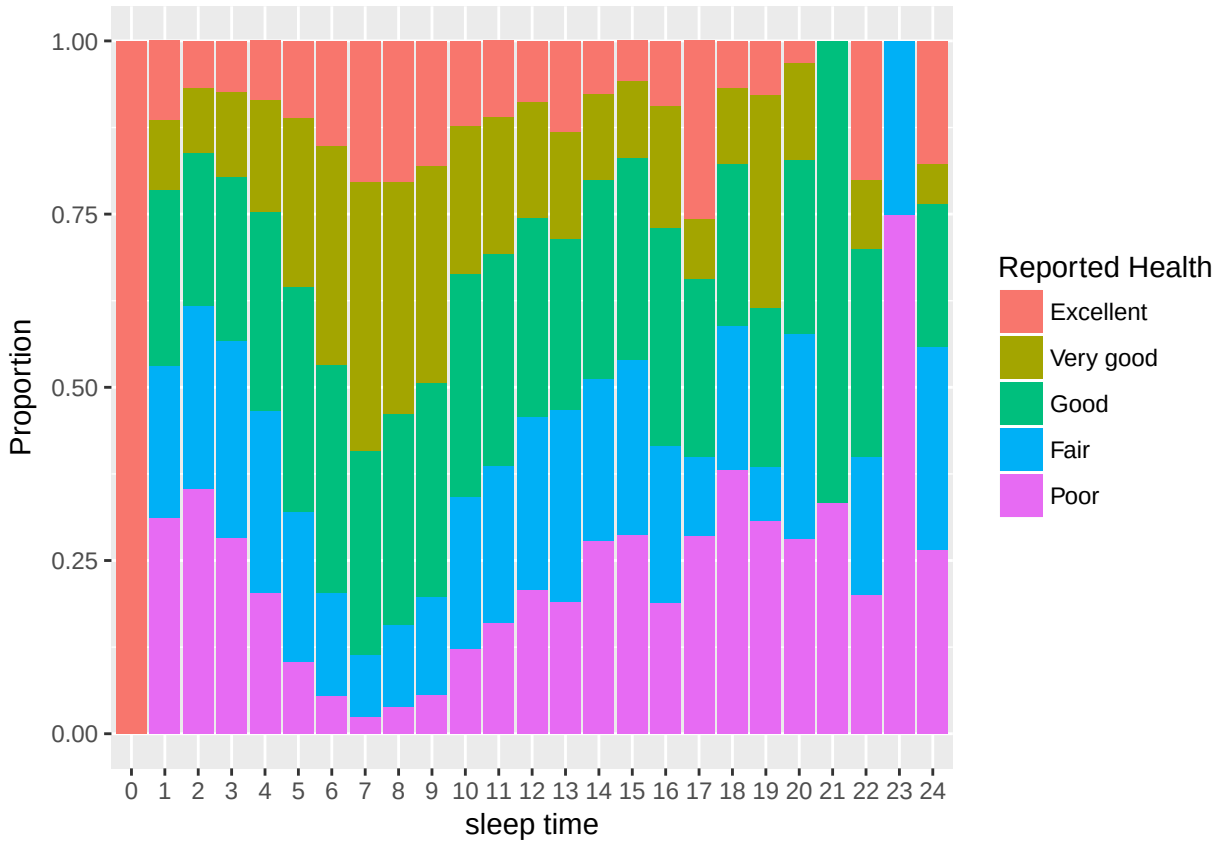
- The mean and stdev of the time people sleep are as above.

```
m <- ggplot(brfss2013, aes(x=1, y=sleep_time)) + geom_boxplot()
m <- m + scale_y_continuous(breaks = round(seq(min(sleep_time,na.rm =TRUE),max(sleep_time,na.rm = TRUE)
m
```



- The box plot shown above gives us an important information regarding outliers.
- When we talk of real world we dont usually hear of people who sleep less than 3 hours or more than 11 hours on an average. This is the reason that there are a lot of outliers in this box plot.
- These outliers indicate that either the people sleeping less than 3 hours and more than 11 hours are very rare, or it may be a data miss-entry or people would have misunderstood the survey question and answered it incorrectly or it can be an answer given in haste.
- In any of the above cases two things are clear, that such people are very rare or they just don't exist, i.e. it is a error or data entry.

```
d1 <- select(brfss2013,genhlth,sleptim1) %>% na.omit()
d1 <- d1 %>% mutate(sleptim1 = as.factor(sleptim1))
c <- ggplot(d1) + aes(x=sleptim1,fill=genhlth) + geom_bar(position = "fill")
c <- c + xlab("sleep time") + ylab("Proportion") + scale_fill_discrete(name="Reported Health")
c
```



- The above graph gives us a very important insight regarding the relationship between average sleep time and general health.
- It is very clear from the graph that the proportion of reported health being excellent near the mean(7.050986) is more than its proportion at other values of average sleep times(x-axis). Also the proportion of poor health near mean hours of sleep is less in proportion than the proportion of poor health at any other average sleep times(x-axis).
- Also the graph follows a specific trend which is:
  - The proportion of excellent health is the least at the edges of the graph and Maximum near the mean.
  - The proportion of poor health is maximum at the edges and minimum at the mean.
  - This tells us that if a person sleeps extremly less or more then needed,the probability of his or her's health being poor would be more, however if he or she sleeps in the range of 6.5 hours to 9.5 hours then the probability of his or her health being excellent would be more. This range also happends to almost coincide with 1 standard deviation from mean range which is[5.58 hr to 8.52 hr]. All this is visible from the graph.
- Conclusion: Thus we can conclude that sleeping properly is indeed useful for a person's health and he or she should sleep atleast forom 6.5 to 9.5 hours and there is a strong correlation between average sleeping time and general health.

Research question 2:

- Here we are converting the long factors of income2 variable to a shorter factor and generating a contingency table.

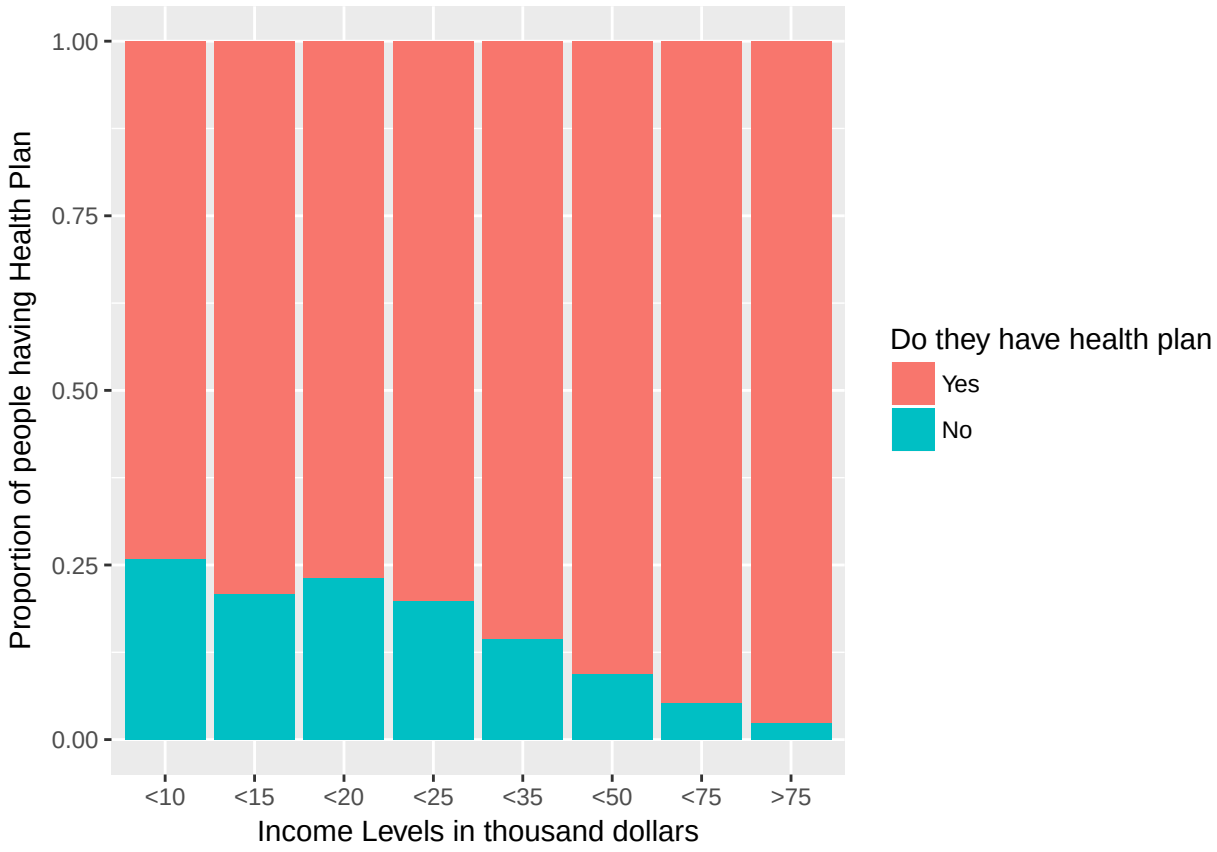
```
income <- brfss2013$income2
income <- as.character(income)
```

```
income[income == "Less than $10,000"] <- "<10"
income[income == "Less than $15,000"] <- "<15"
income[income == "Less than $20,000"] <- "<20"
income[income == "Less than $25,000"] <- "<25"
income[income == "Less than $35,000"] <- "<35"
income[income == "Less than $50,000"] <- "<50"
income[income == "Less than $75,000"] <- "<75"
income[income == "$75,000 or more"] <- ">75"




d3 <- select(brfss2013,genhlth,hlthpln1)
d3 <- cbind(d3,income)
d3 <- d3 %>% na.omit()

prop.table(table(d3$income,d3$hlthpln1),2)
```

```
##
##              Yes         No
##   <10 0.05016492 0.13753994
##   <15 0.05667018 0.11697243
##   <20 0.07167381 0.16959733
##   <25 0.08947427 0.17484500
##   <35 0.11218333 0.14786601
##   <50 0.14942405 0.12287607
##   <75 0.16618884 0.07194397
##   >75 0.30422060 0.05835925
```

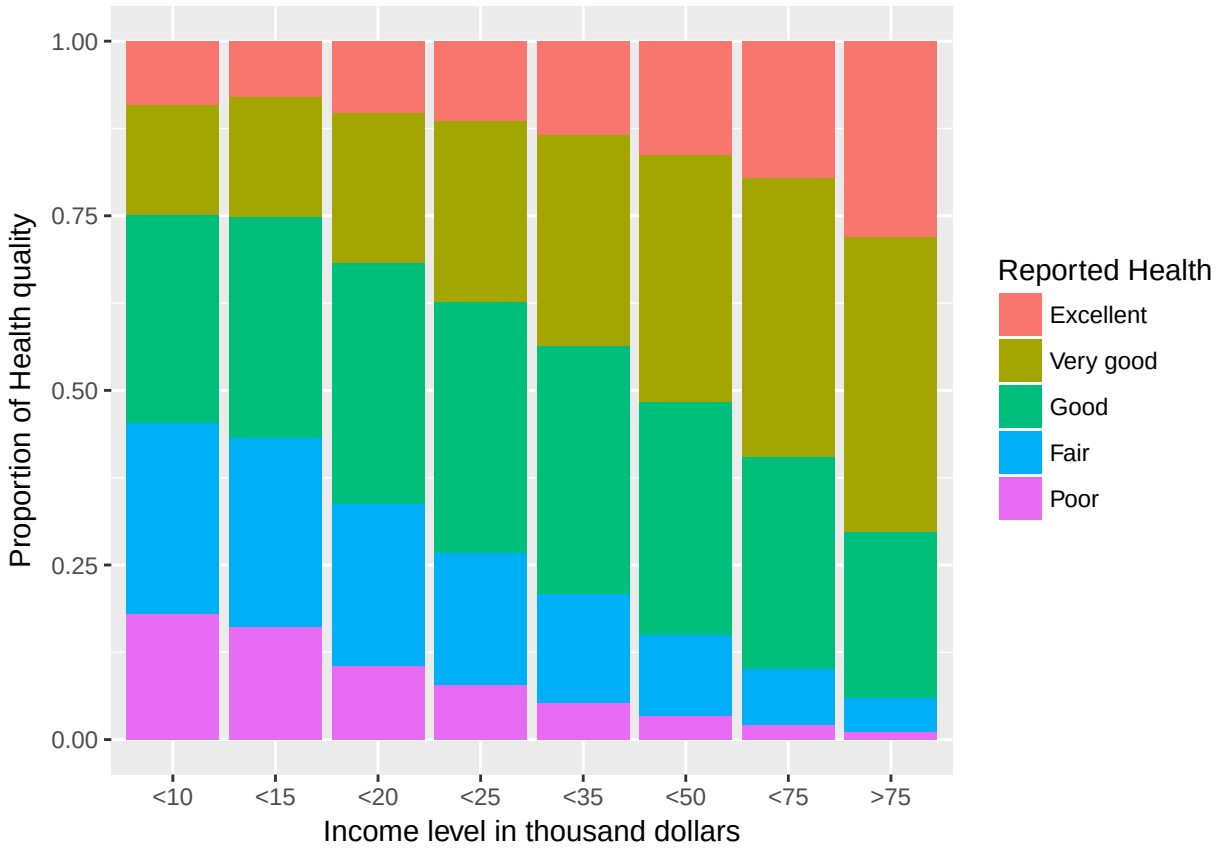
- The above table gives us very important insights about the the income level and the proportion of people from that category that buy the health plans.

```
c <- ggplot(d3) + aes(x=income,fill=hlthpln1) + geom_bar(position = "fill")
c <- c + xlab("Income Levels in thousand dollars") + ylab("Proportion of people having Health Plan") + s
c
```



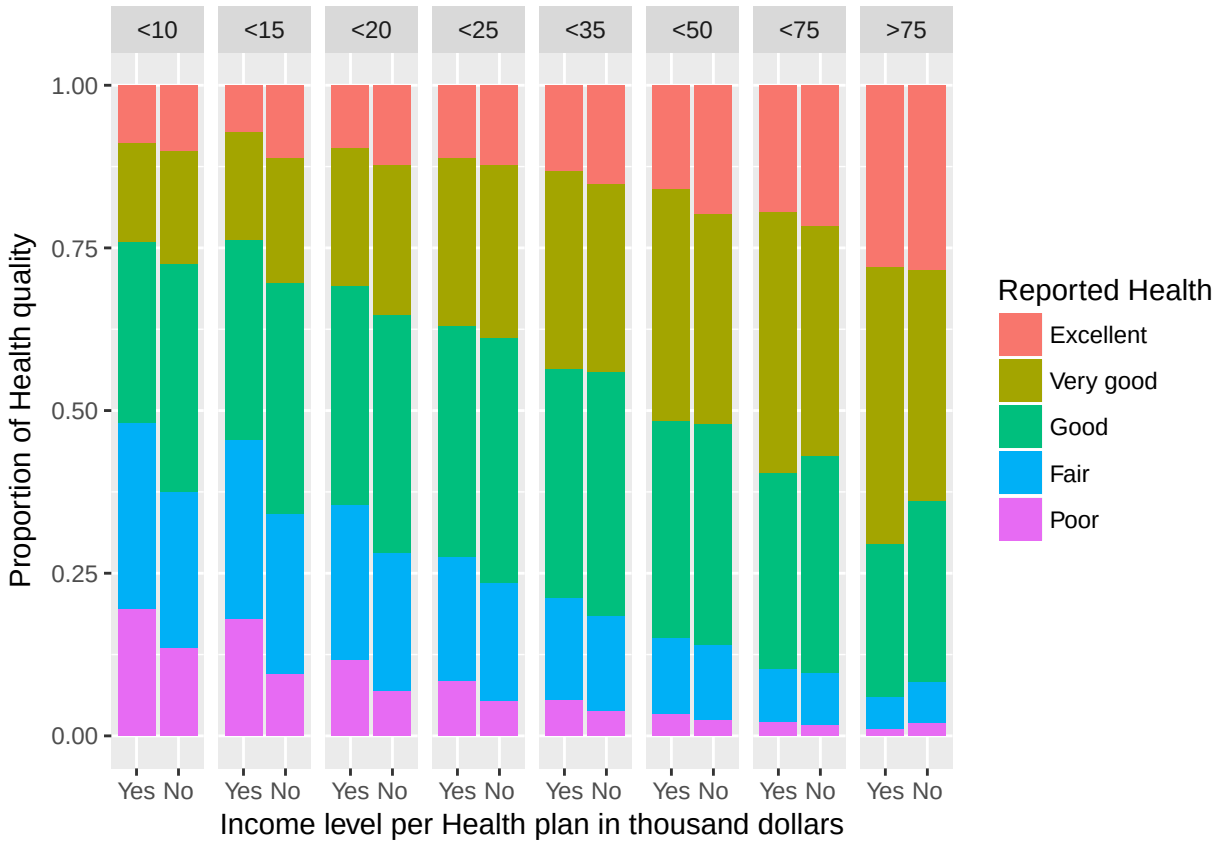
- If you look at this graph you can see that as the income increases , the proportion of people having health plan in that income level also increases.
- This graph gives us the first evidence that the richer you are, the higher is your probability of buying a health plan.

```
c1 <- ggplot(d3) + aes(x=income,fill=genhlth) + geom_bar(position = "fill")
c1 <- c1 + xlab("Income level in thousand dollars") + ylab("Proportion of Health quality") + scale_fill_
c1
```



- The following things can be noticed from this graph:
  - The propertion of people having poor health decreases as the income increases
  - The proportion of people having excellent health increases as the income increases.
- This gives us a hint that income and general health my be directly proportional to each other. However we should look for a confounding variable which might cause such a trend.

```
c1 <- ggplot(d3) + aes(x=hlthpln1,fill=genhlth) + geom_bar(position = "fill") + facet_grid(.~income)
c1 <- c1 + xlab("Income level per Health plan in thousand dollars") + ylab("Proportion of Health quality
c1
```



- This graph shows the same property i.e. the proportion of poor health decreases as the income level increases and the proportion of excellent level increases as the income level increases. There is also a very important property that is mentioned below.
- This is a very interesting graph. Generally we might have expected that if a person has a health plan, his or her's health should be better.However if you look at the lower income levels , even if a person has health plan, his or her proportion of poor health is more than the proportion of poor health when he or she did not have a health plan. This is a bit confusing.
- However we must not forget that an average income below 50,000$ or 25,000$ is still less of an income. In that income the person has to pay taxes, pay his or her rent, take care of the food etc. As a result of this eventhough a person with lower income has a health plan , his chances of ending up in poor health are more as he might be malnutrioned, he may be living in an unhealthy environment etc. There is a confounding variable that is affecting this trend. Also when the income crosses 50,000$ the proportion of poor health with health plan decreases and becomes less than proportion of poor health with no health plan.
- Moreover with higher income as shown in the graph, having a health plan doesn't matter,as the proportion of poor health with and without health plan is neglible. This is because a person with higher income could easily afford medical treaments and has an overall better standard living than a person with lower income.
- Conclusion: Thus we conclude that the health of a person is highly correlated with his inccome level, however having a healthplan does not significantly affect a person's health. Also people with higher income thend to have a higher proportion of health plan, however it has not significant effect on their health. Thus health is highly dependent on income and feebly dependent on health plans.

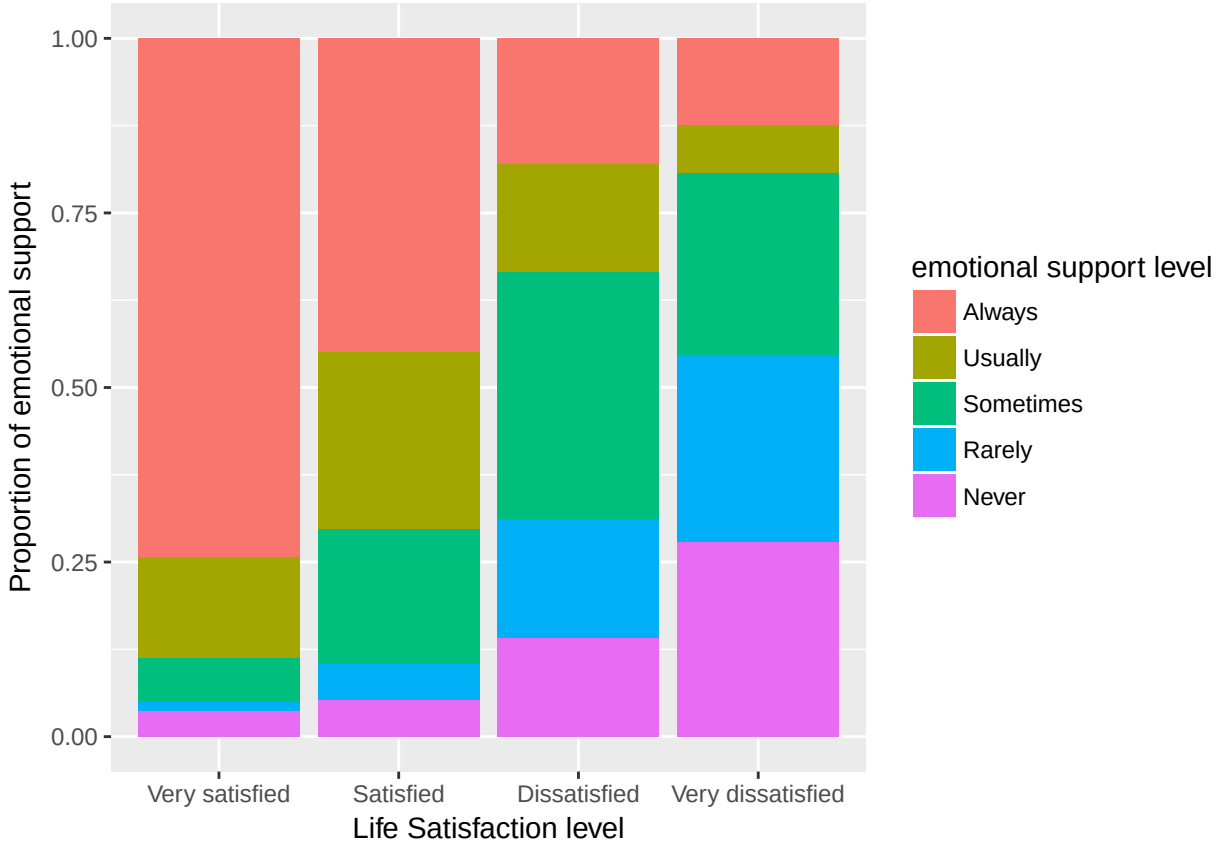
Research quesion 3:

```r
d4 <- select(brfss2013,emtsuprt,lsatisfy,sex) %>% na.omit()

prop.table(table(d4$emtsuprt,d4$lsatisfy),2)
```

```
##
##              Very satisfied  Satisfied Dissatisfied Very dissatisfied
##   Always         0.74292674 0.44829493   0.17935702        0.12422360
##   Usually        0.14427581 0.25382488   0.15566836        0.06832298
##   Sometimes      0.06258197 0.19354839   0.35363790        0.26086957
##   Rarely         0.01405284 0.05142857   0.16920474        0.26708075
##   Never          0.03616264 0.05290323   0.14213198        0.27950311
```

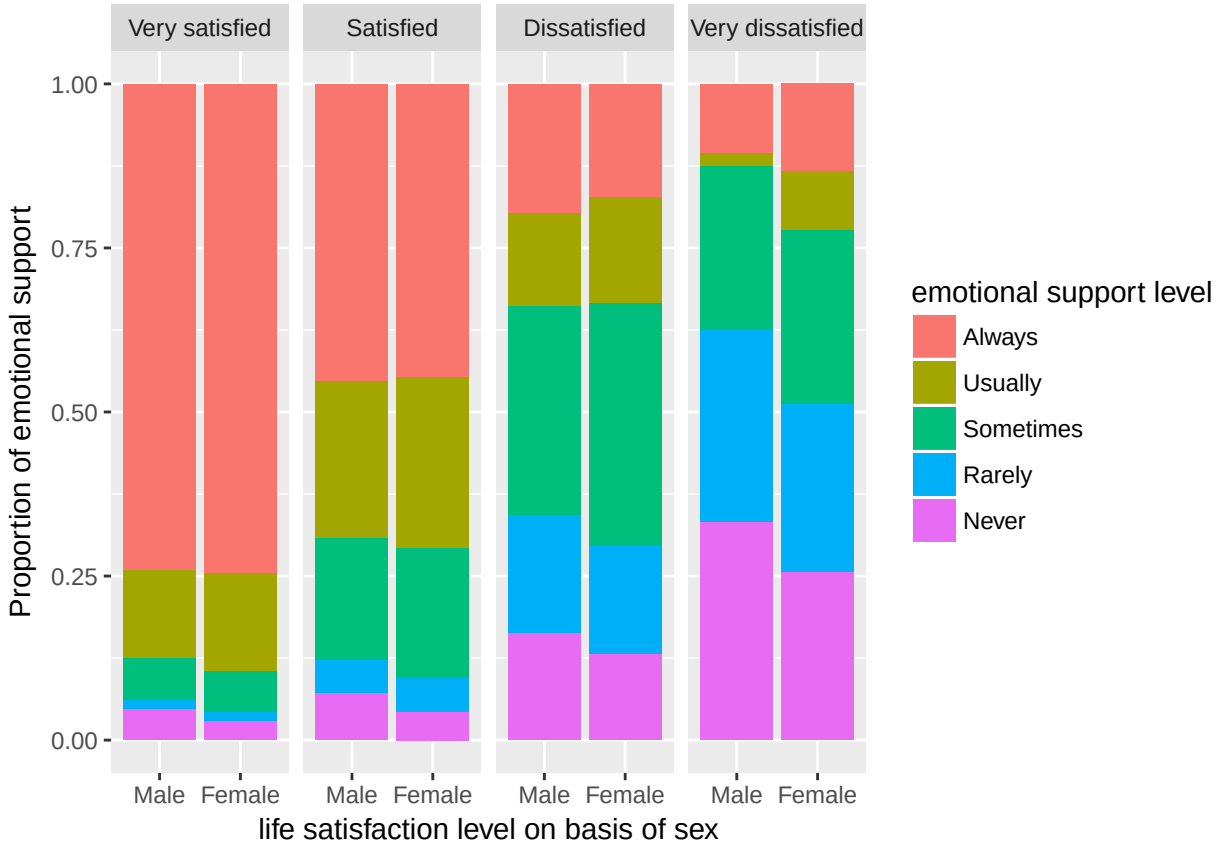
- The above is the contingency table of emotional support received by people and thair satisfaction towards life.

```r
c <- ggplot(d4) + aes(x=lsatisfy,fill=emtsuprt) + geom_bar(position = "fill")
c <- c + xlab("Life Satisfaction level") + ylab("Proportion of emotional support") + scale_fill_discrete
c
```



- Things to be noticed from this graph:
  - Emotional support is maximum when the person is very satisfied with the life, and the satisfaction towards life decreases as the emotional support decreases.Infact when the person is very dissatisfied towards life emotional support is minimum.
  - The people with max satisfaction towards life are almost never devoid of emotional support , however as the as the emotional support starts to decrease , the person starts to get less and less satisfied towards life.
- This gives us an important relation, infact it is a very strong relation between life satisfaction and emotional support. i.e. Life satisfaction is directly propertional to emotional support.

```
c1 <- ggplot(d4) + aes(x=sex,fill=emtsuprt) + geom_bar(position = "fill") + facet_grid(.~lsatisfy)
c1 <- c1 + xlab("life satisfaction level on basis of sex") + ylab("Proportion of emotional support") + s
c1
```



- The above graph is very similar to the one above it, the only diffference is that now we are considering the effects of emotional support on two sexes, i.e. Male and female.
- Also there is not much change in the proportion of emotional support received by male and female.
- The graph also follows the same trend, i.e. life satisfaction increases as the emotional support increases and vice versa for both the sexes.
- Conclusion: Hence we conclude that, irrespective of the sex of a person, emotional support is necessary for anyone to thrive in life. The more you get that emotional support the more you will be satisfied towards life, because afterall humans are social animals.