Exploring the BRFSS data

Setup

Load packages

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

Load data

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

Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS2013) 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. Some questions ask the respondent information about actions and events happened in the past 30 days or even before, and some people might not remember accurately what they have done such a long time before. Another issue is that the interview are contacted over the phone and participation is voluntary, therefore there is no random assignment and also people who do not own a telephone (landline and/or mobile) or do not live in a private residence are excluded from the study.
A random sample from each state has been selected, therefore this is a **stratified random sample**. This is an **observational study**, because it is asking respondent about their past actions, it is not an experimental study as there has not been any random assignment. Therefore we can infer from the observational nature of the study and the fact that participants were not randomly assigned to groups, that the study's results are **generalizable**.
We can further infer that the results of the study are **non-causal** and only correlation statements can be made using the results.

Part 2: Research questions

Research question 1:

Is there a correlation between the number of hours worked in a week and the income level of the respondent? Does this correlation change based on the sex of the respondent?

Research question 2:

Is there a correlation between the number of hours a respondent sleeps everyday and their blood pressure? Are people who work more number of hours in a week more affected?

Research question 3:

Do people who have a higher income have a greater or lesser chance of having a depressive disorder. Does this correlation change when they have more children in the household i.e dependents.

Part 3: Exploratory data analysis

Research question 1:

Is there a correlation between the number of hours in a week a respondent works and their income level? Does this relationship change based on the gender of the respondent.

Do people who earn less money or have daily wage jobs tend to work more hours in a week for sustenance. Or maybe individuals with higher profile jobs have to spend more time working due to large amounts of responsibilities and pressure to deliver results and innovations. Although often higher paying jobs may allow individuals to have a better work life balance. We also try to determine if there is a noticeable difference in the relationship of work to earnings between men and women.

The variables used to answer this question are as follows:
1. scntwrk1 - How many hours per week do you work (continuous variable [0-98])
2. income2 - Income level (Ordinal variable)
3. sex - Respondents sex (Binary variable)

We clean the NA variables from scntwrk1, income2, and sex columns. The result is then stored in Dataframe2 variable.

```
Dataframe2 = brfss2013 %>% filter(!(is.na(income2)), !(is.na(scntwrk1)), !
(is.na(sex)))
```

We change the values of the income variable(income2) to values with shorter variables that are easier to read and work with. The values are also more convenient while plotting as they take us less unnecessary space.
We then observe the summary statistics of the income variable.

```
levels(Dataframe2$income2) = c("< $10k", "< $15k", "< $20k","< $25k", "<
$35k", "< $50k", "< $75k", ">= $75k")
Dataframe2 %>% group_by(income2) %>% summarise(count = n())
```

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

```
## # A tibble: 8 x 2
##   income2 count
##   <fct>   <int>
## 1 < $10k    500
## 2 < $15k    692
## 3 < $20k   1396
## 4 < $25k   1988
## 5 < $35k   2802
## 6 < $50k   4040
## 7 < $75k   5418
## 8 >= $75k 12757
```

We then calculate the mean, median and standard deviation for the number of hours respondent works in a week(scntwrk1) for each income category(income2).

```
work_and_income = Dataframe2 %>% group_by(income2) %>% summarise(mean_work
= mean(scntwrk1), median_work = median(scntwrk1), sd_work = sd(scntwrk1))
```

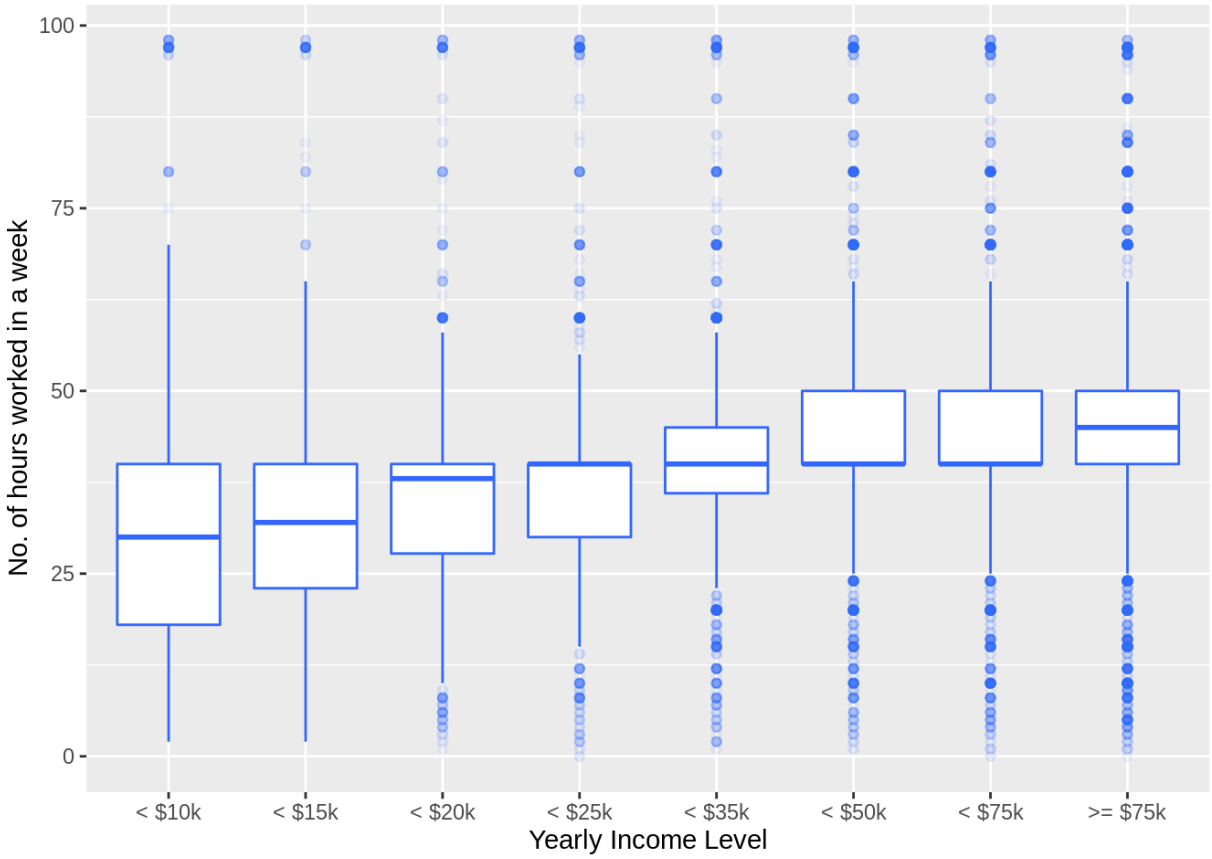
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
work_and_income
```

```
## # A tibble: 8 x 4
##   income2 mean_work median_work sd_work
##   <fct>       <dbl>       <dbl>   <dbl>
## 1 < $10k       35.2          30    25.7
## 2 < $15k       35.4          32    20.9
## 3 < $20k       36.5          38    17.2
## 4 < $25k       39.1          40    16.9
## 5 < $35k       41.4          40    15.6
## 6 < $50k       43.2          40    15.0
## 7 < $75k       43.8          40    14.1
## 8 >= $75k      45.2          45    14.0
```

The weekly work hours and income level are then represented graphically in the form of a boxplot.

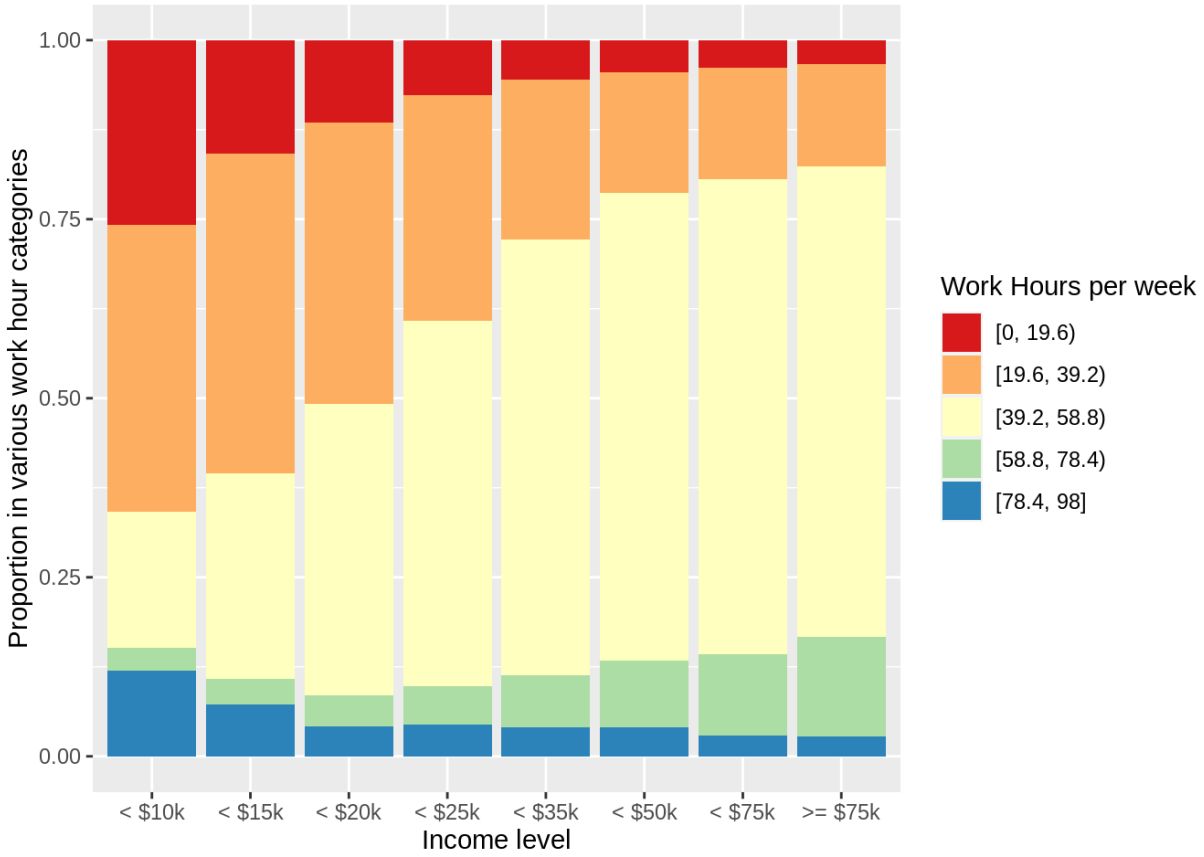
```
income_vs_work_bp = ggplot(Dataframe2, aes(x = income2, y = scntwrk1)) +
geom_boxplot(fill = "white", colour = "#3366FF", outlier.colour =
"#3366FF", outlier.alpha = 0.05) + xlab("Yearly Income Level") +ylab("No.
of hours worked in a week")
income_vs_work_bp
```



It is evident from the above table and boxplot representation that, lower income respondents tend to spend less number of hours working in a week and as the respondents income increases, they spend more number of hours working in a week.

We then divide the weekly work hours(scntwrk1) into discrete intervals to better represent them graphically.
The following is a bar graph representation between the income levels(income2) and Weekly work hours.

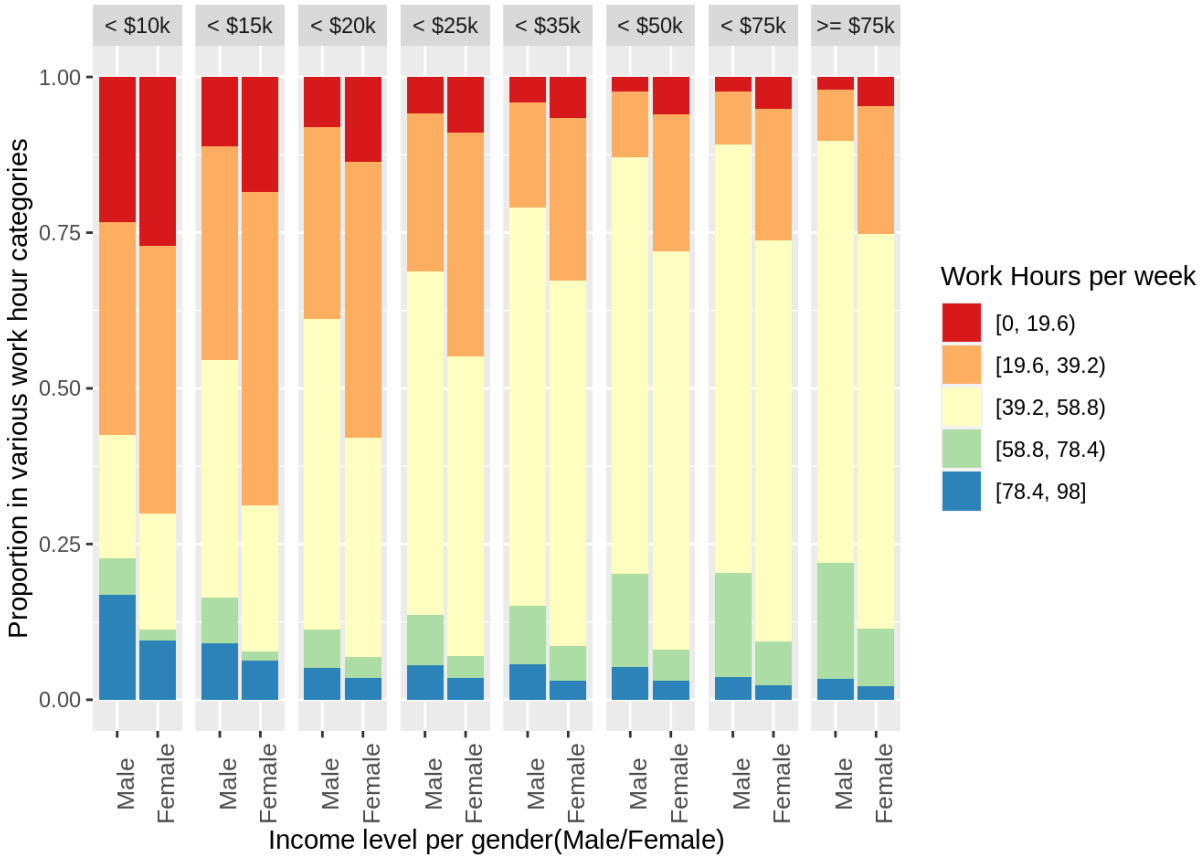
```
Dataframe2[, "work_hours"] = bin_data(Dataframe2$scntwrk1, bins = 5,
binType = "explicit")

bar_plot = ggplot(Dataframe2, aes(x=income2, fill = work_hours)) +
geom_bar(position = "fill") + xlab("Income level") + ylab("Proportion in
various work hour categories") + scale_fill_brewer(name = "Work Hours per
week", palette = "Spectral")
bar_plot
```



Another interesting observation we make is that respondents that have higher incomes tend to have more reasonable working hours i.e the spread of work hours is less, and large number of respondents have weekly work hours around the median number of hours. Whereas, respondents with lower income tend to work more unreasonable and extreme number of hours in a week i.e a large proportion of them tend to work less than 20 hrs a week or more than 80 hours a week. This leads to the conclusion that higher income respondents tend to have much better work life balance.

We then add another dimension to the previous graphical representation i.e the gender of the respondent, in an attempt to determine if the weekly work hours and income level correlation changes.

```
work_income_sex_barplot = ggplot(Dataframe2, aes(x = sex, fill =
work_hours)) + geom_bar(position = "fill") + facet_grid(.~income2) +
xlab("Income level per gender(Male/Female)") + ylab("Proportion in various
work hour categories") + scale_fill_brewer(name="Work Hours per week",
palette = "Spectral") + theme(axis.text.x = element_text(angle = 90, size
= 10))
work_income_sex_barplot
```



From the above graph, we can infer that in each income category female respondents tend to work less number of hours in a week. At each income level, a larger proportion of females work less than 40 hrs a week as compared to the males at the same income level. Also a lower proportion of females work more than 60 hrs a week as compared to men at the same income level. A likely cause for this is the disproportionate distribution of household responsibilities i.e females are often expected to fulfill household responsibilities regardless of whether they work or not.

Research question 2:

Is there a correlation between the number of hours a respondent sleeps everyday and their blood pressure? Are people who work more number of hours in a week more affected?
The significance of this problem is to figure out if people who work more hours a weeks tend to sleep less. And if these habits of more work and less sleep have an effect on their health, specifically on their blood pressure.

The variables used to answer this question are:
1. scntwrk1 - How many hours per week do you work (continuous variable [0-96])
2. sleptim1 - How much time do you sleep (continuous variable [1-24])
3. bphigh4 - Ever told blood pressure high (categorical variable)

Removing NA values from all columns scntwrk1, sleptim1, bphigh4

```
Dataframe1 = brfss2013 %>% filter(!is.na(scntwrk1) , !is.na(sleptim1), !
is.na(bphigh4))
```

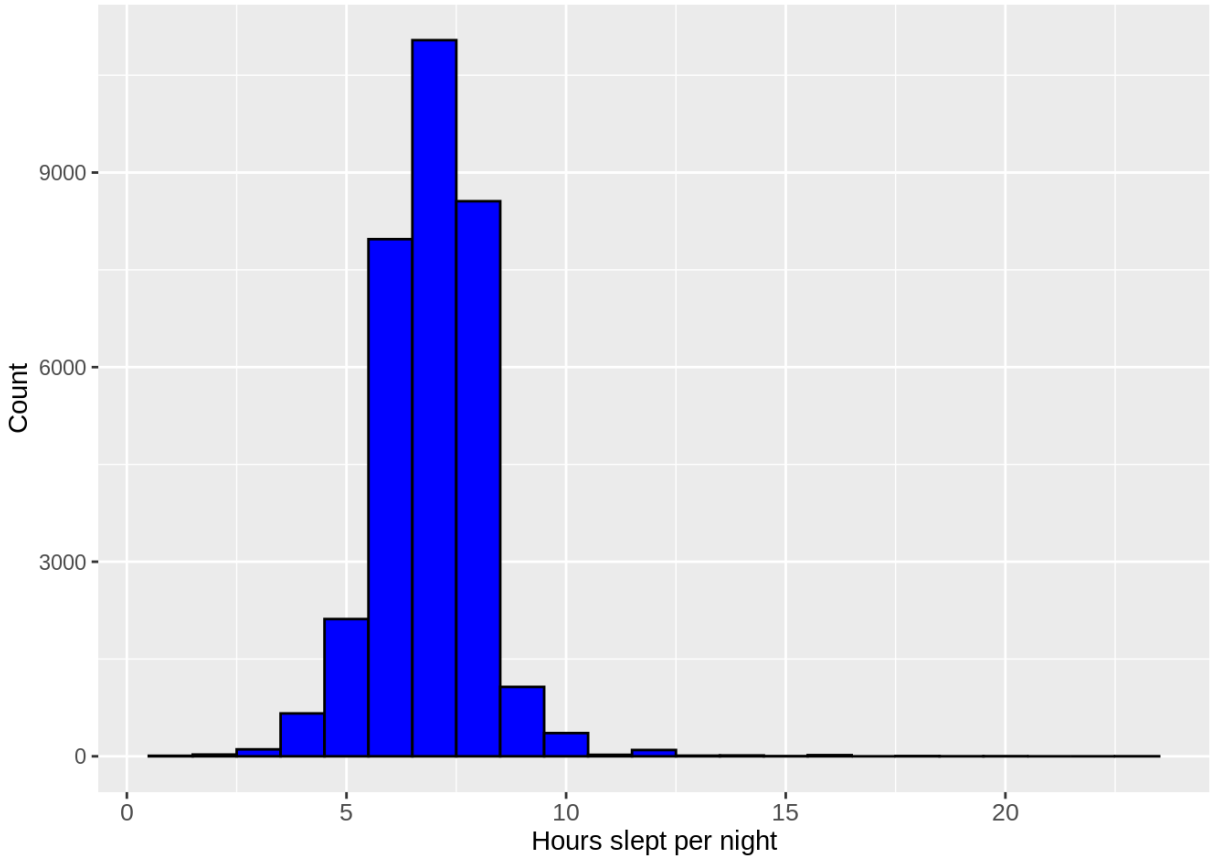
Calculating summary statistics of all the columns to get an idea of what kind of values they have. For average number of hours slept per day in the last week:

```
Dataframe1 %>% summarise(mean_sleep = mean(sleptim1), median_sleep =
median(sleptim1), sd_sleep = sd(sleptim1), min_sleep = min(sleptim1),
max_sleep = max(sleptim1))
```

```
##   mean_sleep median_sleep sd_sleep min_sleep max_sleep
## 1    6.93721            7 1.218642         1        23
```

Now create a graphical representation of the respondents on number of hours slept

```
Sleep_plot = ggplot(data = Dataframe1, aes(x = sleptim1)) +
geom_histogram(binwidth = 1, color = "black", fill = "blue") + xlab("Hours
slept per night") + ylab("Count")+theme(axis.text.x = element_text(size =
10))
Sleep_plot
```



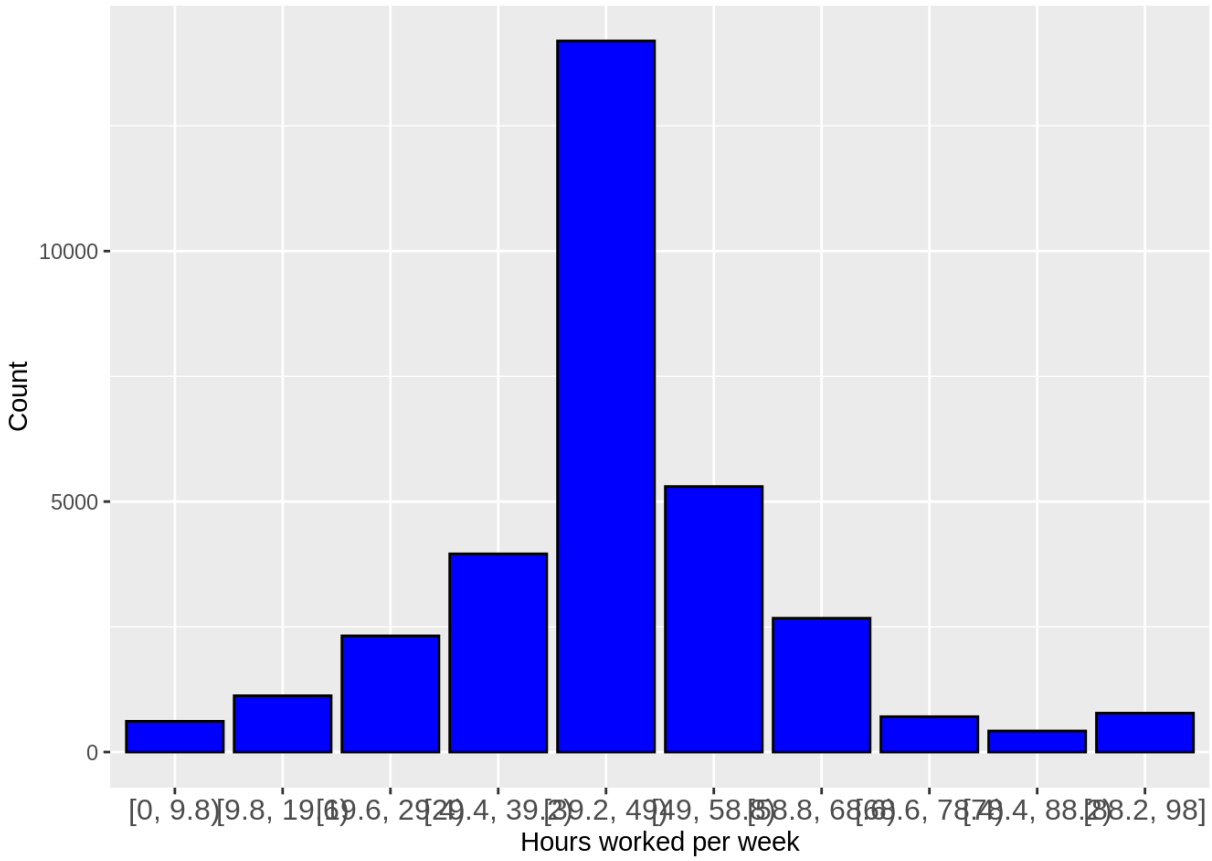
From the above graph we can infer that the majority of the respondents sleep between 4-11 hrs in a day. *** For how many hours respondents work per week:

```
Dataframe1 %>% summarise(mean_work = mean(scntwrk1), median_work =
median(scntwrk1), sd_work = sd(scntwrk1), min_work = min(scntwrk1),
max_work = max(scntwrk1))
```

```
##   mean_work median_work  sd_work min_work max_work
## 1  42.98501          40 15.73863        0       98
```

For the hours worked in a week column, we convert the given continuous values into equal sized discreet groups and represent graphically using a bar graph.

```
library(mltools)
Dataframe1[, "work_hours"] = bin_data(Dataframe1$scntwrk1, bins = 10,
binType = "explicit")

work_plot = ggplot(data = Dataframe1, aes(x = work_hours)) +
geom_bar(color = "black", fill = "blue") + xlab("Hours worked per week") +
ylab("Count")+theme(axis.text.x = element_text(size = 12))
work_plot
```



We now calculate the summary statistics for hours slept per night for each group in the work_hours column.

```
work_and_sleep = Dataframe1 %>% group_by(time_worked =
as.factor(work_hours)) %>% summarise(mean_sleep = mean(sleptim1),
median_sleep = median(sleptim1), sd_sleep = sd(sleptim1), count = n())
```

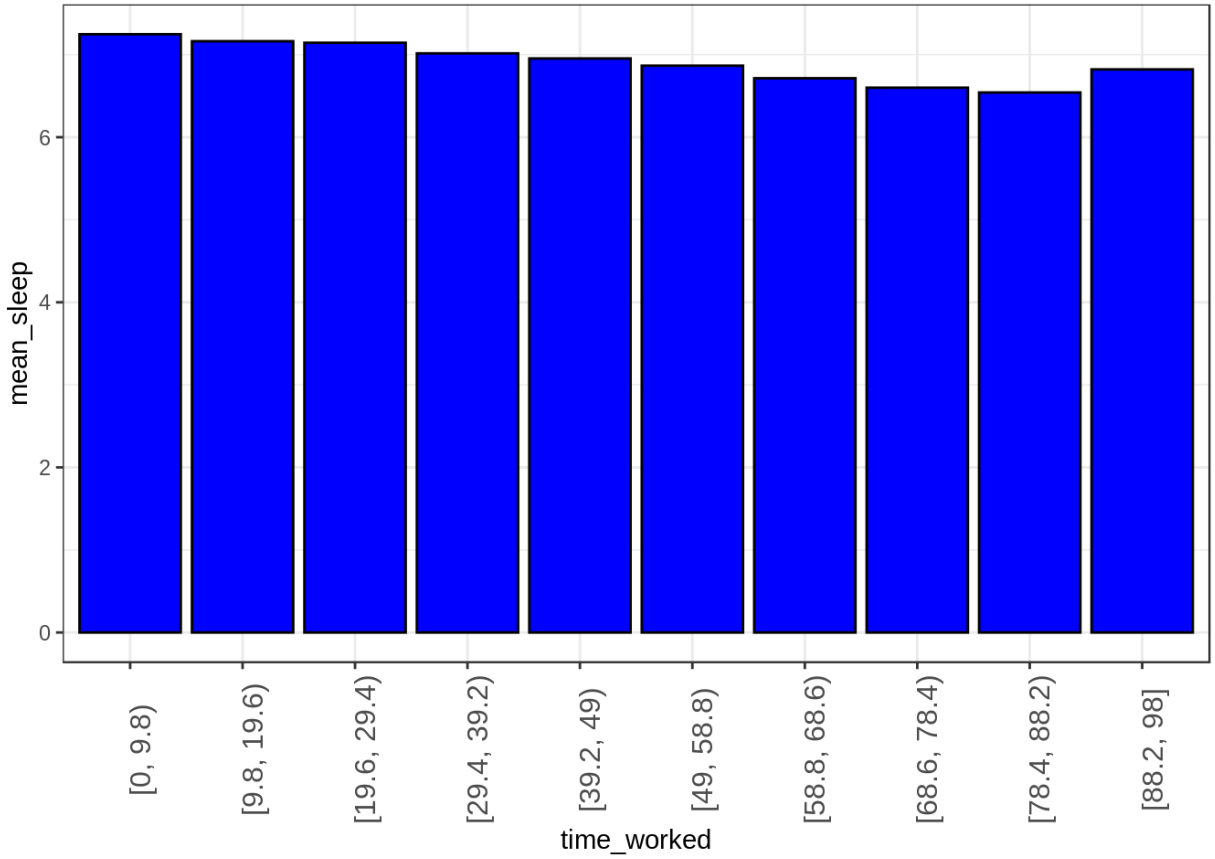
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
work_and_sleep
```

```
## # A tibble: 10 x 5
##    time_worked  mean_sleep median_sleep sd_sleep count
##    <ord>             <dbl>        <dbl>    <dbl> <int>
##  1 [0, 9.8)           7.25            7     1.33   615
##  2 [9.8, 19.6)        7.16            7     1.44  1125
##  3 [19.6, 29.4)       7.14            7     1.34  2320
##  4 [29.4, 39.2)       7.02            7     1.28  3957
##  5 [39.2, 49)         6.95            7     1.17 14195
##  6 [49, 58.8)         6.87            7     1.06  5301
##  7 [58.8, 68.6)       6.71            7     1.15  2672
##  8 [68.6, 78.4)       6.60            7     1.40   707
##  9 [78.4, 88.2)       6.54            6     1.49   421
## 10 [88.2, 98]         6.82            7     1.55   778
```

We firstly plot a bar graph with time_worked along the x-axis and mean of hours slept every night along the y-axis.

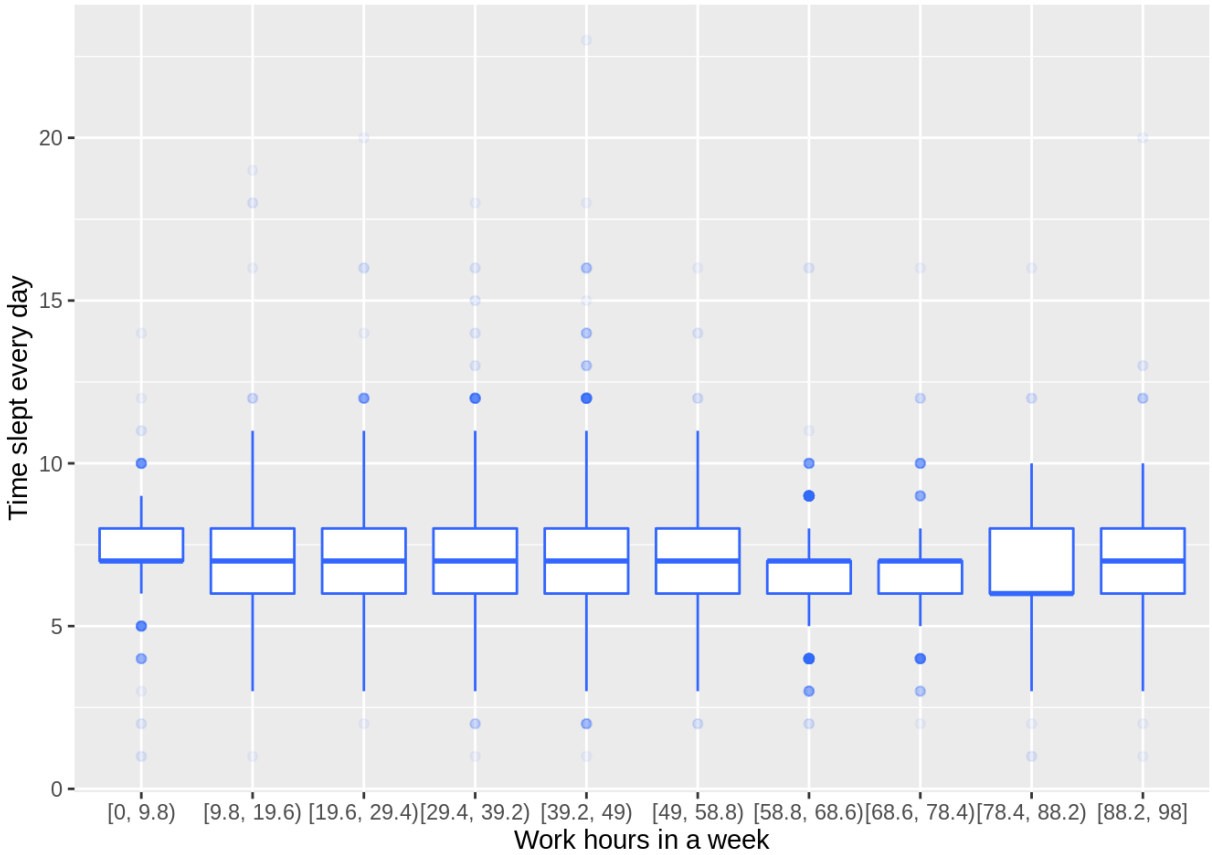
```
work_and_sleep_plot = ggplot(data = work_and_sleep, aes(x = time_worked, y
= mean_sleep)) + geom_bar(stat = 'identity', color="black", fill="blue") +
theme_bw() + theme(axis.text.x = element_text(angle = 90, size = 12))
work_and_sleep_plot
```



On observation of the above graph, we can infer that apparent decrease in the mean time slept everyday by the respondent as the number of work hours in a week increases.
But the above graph takes the mean sleep time which may not be an accurate measure as mean values are quite affected by extreme values. Despite the fact that most respondents sleep between 4-11 hrs everyday, there are respondents in each work hour category that sleep as high as 23 hrs or as low as 1 hr in a day.
For this reason, we decide that the inferences from the above graph may not be accurate.

We can also represent the summary statistics in the form of a boxplot which gives us a more accurate visual representation of sleep hours with respect to Working hours.

```
work_vs_sleep_boxplot = ggplot(Dataframe1, aes(x = work_hours, y =
sleptim1)) + geom_boxplot(fill="white", colour = "#3366FF", outlier.color
= "#3366FF", outlier.alpha = 0.05) + xlab("Work hours in a week") +
ylab("Time slept every day")
work_vs_sleep_boxplot
```



From the boxplot we get values such as median, quartiles and inter quartile range(IQR) which are a more accurate representation of the data as these quantities are not affected by extreme values. As we can see in the above plot, there is no significant difference between the median hours slept or IQR for each weekly work hour category.
Hence we conclude that there is no difference correlation between the number of hours worked in a week and the amount of time slept every day.

High blood pressure column has 4 distinct possible values i.e Yes, No, Told borderline or pre-hypertensive, Yes but female told only during pregnancy. We are interested in only yes and no answers. We then tabulate how many respondents ever reported having high blood

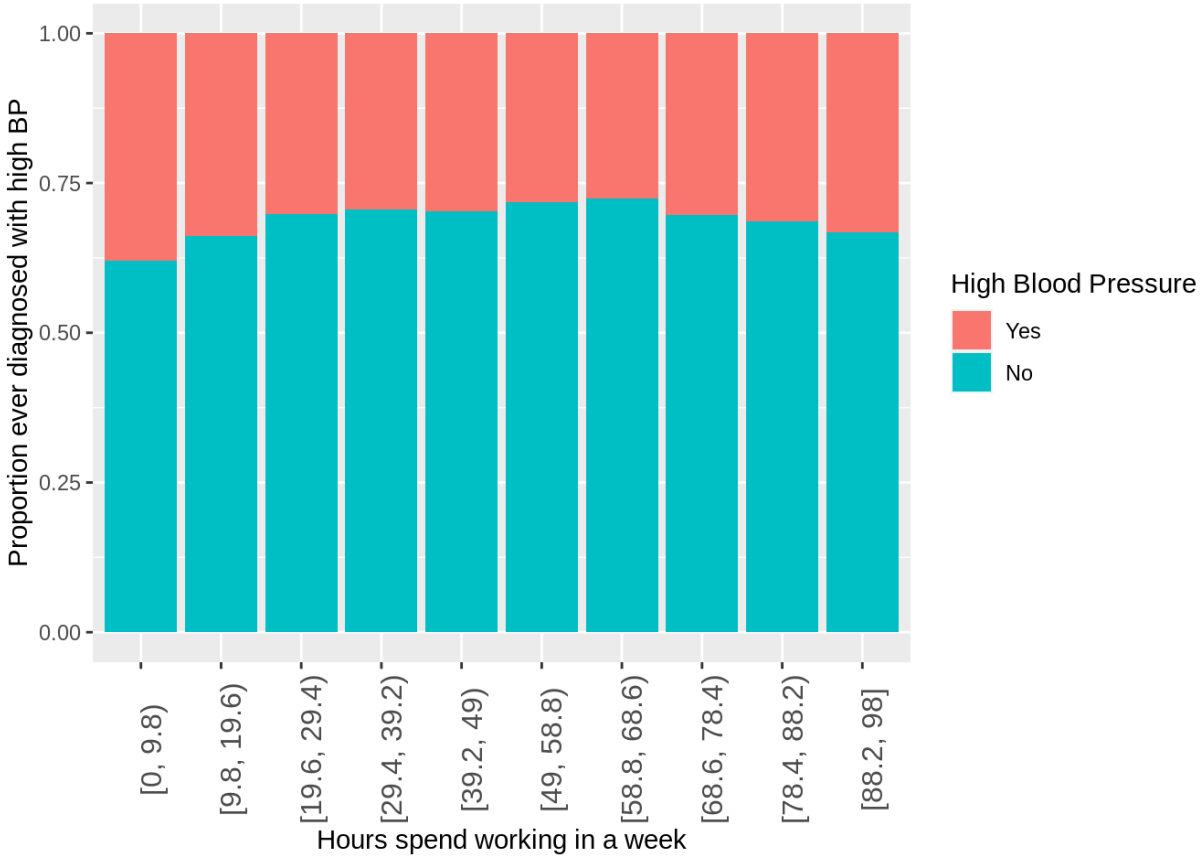
```
Dataframe1 = Dataframe1 %>% filter(bphigh4 == "Yes" | bphigh4 == "No")
Dataframe1 %>% group_by(bphigh4) %>% summarise(count = n())
```

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

```
## # A tibble: 2 x 2
##   bphigh4 count
##   <fct>   <int>
## 1 Yes      9342
## 2 No      22108
```

To Plot the relation between number of hours spent working in the week and its affect on respondents blood pressure

```
work_vs_bp = ggplot(data = Dataframe1, aes(x = work_hours, fill =
bphigh4)) + geom_bar(position="fill") + scale_fill_discrete(name = "High
Blood Pressure") + xlab("Hours spend working in a week") +
ylab("Proportion ever diagnosed with high BP") + theme(axis.text.x =
element_text(angle = 90, size = 12))
work_vs_bp
```



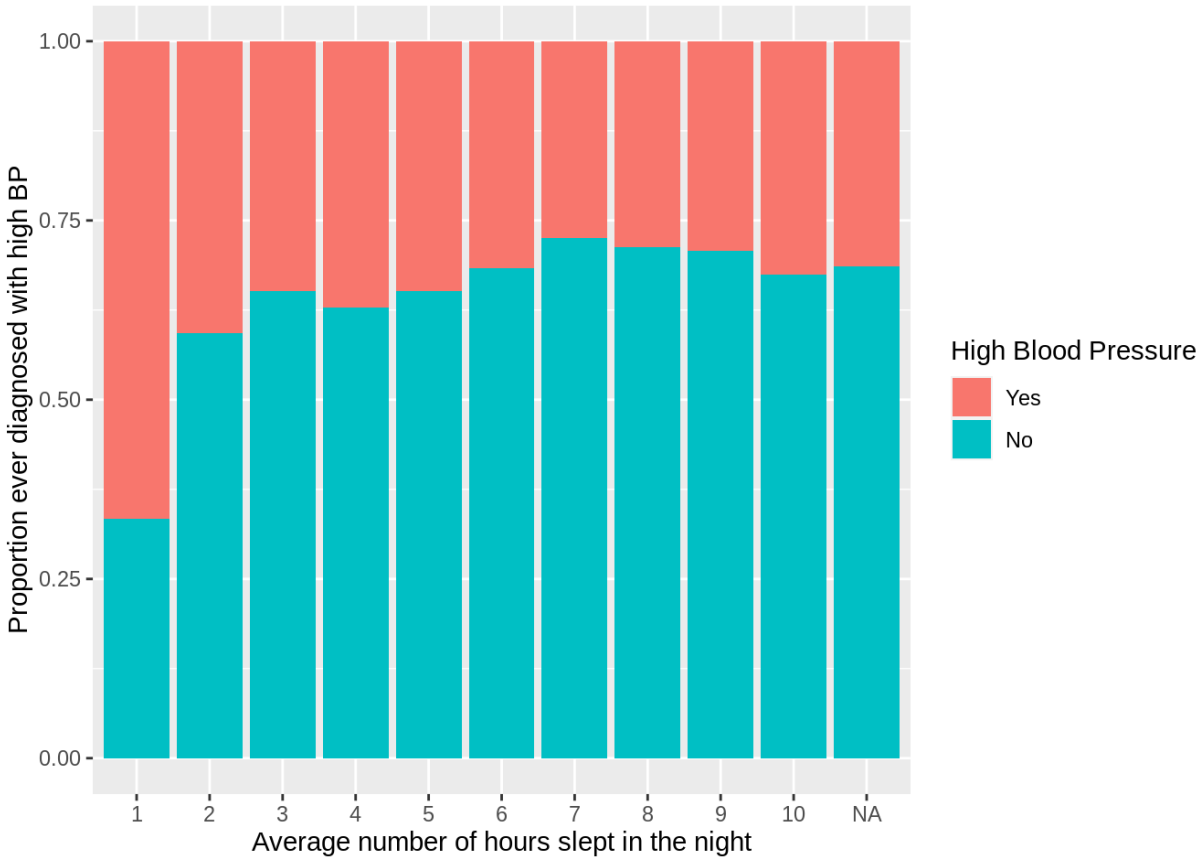
From the above graphical representation, we infer that there doesn't seem to be any obvious correlation between the proportion of respondents with high blood pressure and hours they spend working in a week.
We can see however that the proportion of respondents with high blood pressure seems to be higher among extreme work hour categories i.e less than 10hrs a week and more than 88 hrs in a week.

We now split the sleep hours into 11 categories are shown below and most respondents lie in 1-11 hour categories.

```
Dataframe1$sleptim1 = factor(Dataframe1$sleptim1, levels = c("1", "2",
"3", "4", "5", "6", "7", "8", "9", "10", "11+"))
```

We now plot the relation between the number of hours slept per day and its affect on respondents blood pressure

```
sleep_vs_bp = ggplot(data = Dataframe1, aes(x = sleptim1, fill = bphigh4))
+ geom_bar(position="fill") + scale_fill_discrete(name = "High Blood
Pressure") + xlab("Average number of hours slept in the night") +
ylab("Proportion ever diagnosed with high BP")
sleep_vs_bp
```



In the above graph we can see that almost 70% of the respondents that sleep only 1 hr a day have high blood pressure. The proportion of respondents with high blood pressure tends to be lower when they have more reasonable sleep hours.

Research question 3:

Do people who have a higher income have a greater or lesser chance of having a depressive disorder? Is this correlation affected if the respondent has more children in the household i.e dependents?

We often come across headlines in the news stating so and so rich and famous person is fighting depression or has committed suicide. This leads us to believe the rich are often depressed and it seems as though this number is quite comparable if not more than the number of less economically well off people who suffer from a depressive disorder. The significance of this question is to answer whether this is just a perception created due to news reports as news reports tend to only talk about the rich and famous as they make better headlines.

The reason for exploring the correlation between income level and depression correlation is to see if having an abundance of material things in ones life has an impact on psychological life.

The variables used to answer this question are:
1. income2 - Income level (Ordinal variable)
2. addepev2 - Ever told you have a depressive disorder (Binary variable)
3. children - Number of children in the household (Discrete variable [0-20])

We first evaluate the above variables by summarizing the various values present in each column. We summarize the values present in the income level column(income2).

```
brfss2013 %>% group_by(income2) %>% summarise(count = n())
```

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

```
## # A tibble: 9 x 2
##   income2            count
##   <fct>              <int>
## 1 Less than $10,000  25441
## 2 Less than $15,000  26794
## 3 Less than $20,000  34873
## 4 Less than $25,000  41732
## 5 Less than $35,000  48867
## 6 Less than $50,000  61509
## 7 Less than $75,000  65231
## 8 $75,000 or more   115902
## 9 <NA>               71426
```

We summarize the values present in the depressive disorder column(addepev2).

```
brfss2013 %>% group_by(addepev2) %>% summarise(count = n())
```

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

```
## # A tibble: 3 x 2
##   addepev2  count
##   <fct>     <int>
## 1 Yes       95779
## 2 No       393707
## 3 <NA>       2289
```

We summarize the values present in the column that consists of number of children in the household(children).

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

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

```
## # A tibble: 21 x 2
##    children  count
##       <int>  <int>
##  1        0 359478
##  2        1  53208
##  3        2  46682
##  4        3  19828
##  5        4   6924
##  6        5   2099
##  7        6    827
##  8        7    240
##  9        8    103
## 10        9     45
## # … with 11 more rows
```

To retrieve the columns for Income level(income2), Presence of depressive disorder(addepev2) and number of children in the household(children), and remove all NA values.

```
Dataframe3 = brfss2013 %>% filter(!is.na(income2), !is.na(addepev2), !
is.na(children)) %>% select(income2, addepev2, children)
```

We change the values of the income variable(income2) to values with shorter variables that are easier to read and work with. The values are also more convenient while plotting as they take us less unnecessary space.
We then observe the summary statistics of the income variable.

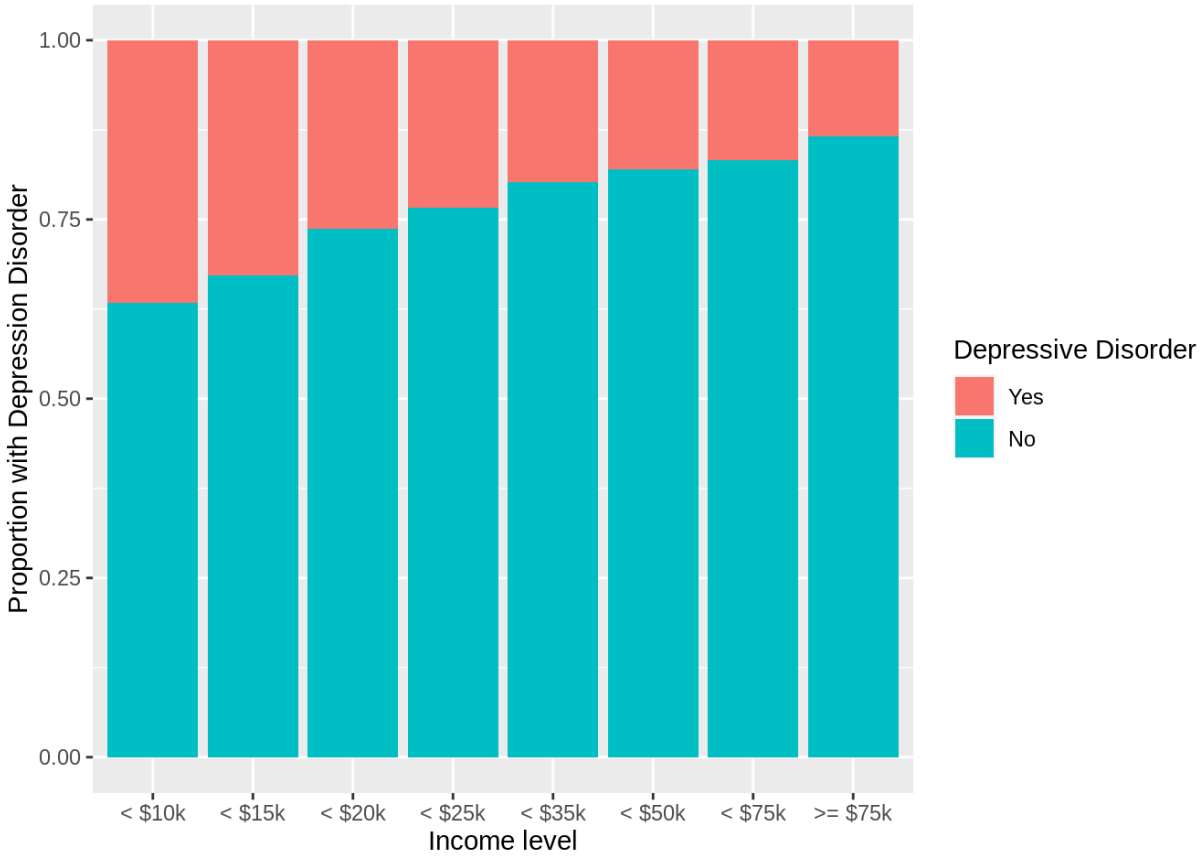
```
levels(Dataframe3$income2) = c("< $10k", "< $15k", "< $20k","< $25k", "<
$35k", "< $50k", "< $75k", ">= $75k")
Dataframe3 %>% group_by(income2) %>% summarise(count = n())
```

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

```
## # A tibble: 8 x 2
##   income2  count
##   <fct>    <int>
## 1 < $10k   25184
## 2 < $15k   26592
## 3 < $20k   34669
## 4 < $25k   41455
## 5 < $35k   48621
## 6 < $50k   61234
## 7 < $75k   64976
## 8 >= $75k 115471
```

To graphically represent the relation between Income level and Proportion with Depression Disorder

```
income_vs_depression_plot = ggplot(Dataframe3, aes(x=income2,
fill=addepev2)) + geom_bar(position="fill") + xlab("Income level") +
ylab("Proportion with Depression Disorder") + scale_fill_discrete(name =
"Depressive Disorder")
income_vs_depression_plot
```



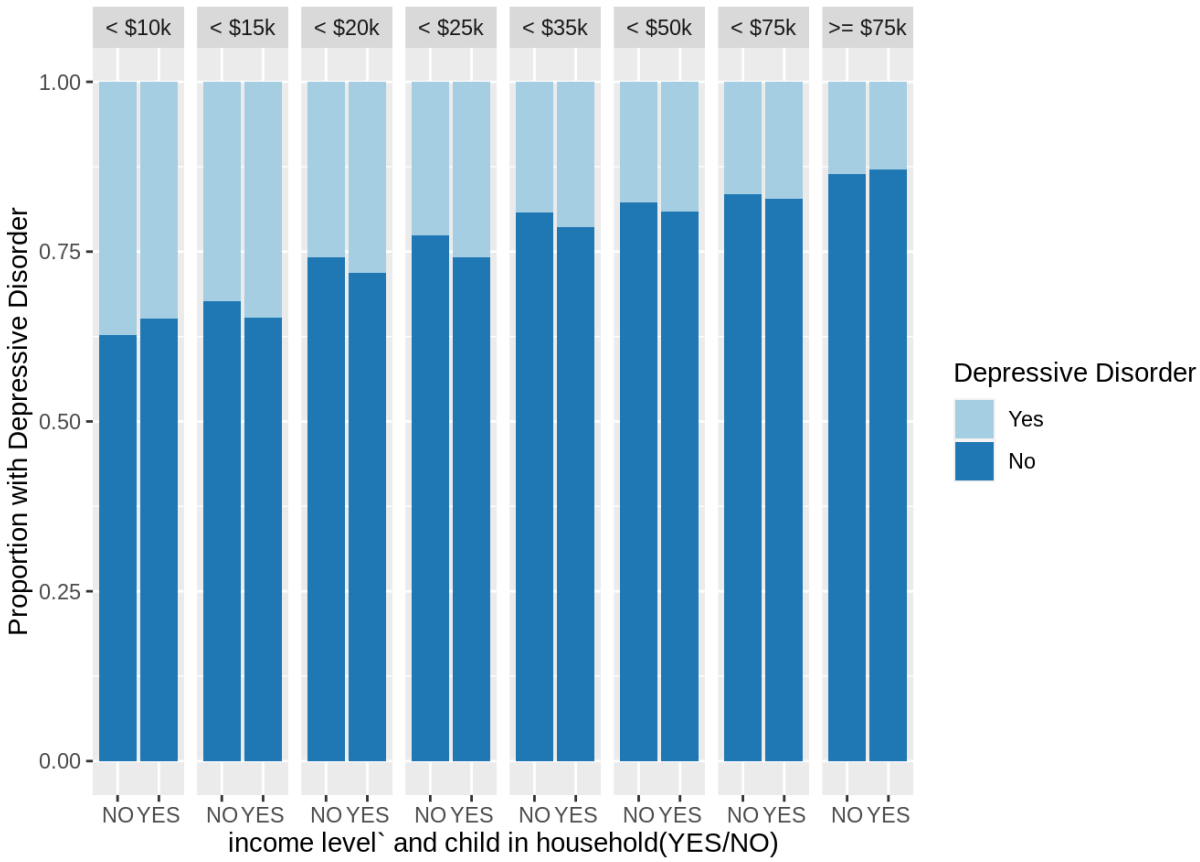
In this graphical representation, when we consider the study participant according to their income level, the proportion of those who have ever been diagnosed with a depressive disorder is higher among those with a lower income.

We now add a new variable to the binary variable(has_child) to the dataframe in which if the respondent has a child then has_child = “YES” and if the respondent doesn’t have a child then has_child = “NO”.

```
Dataframe3 = Dataframe3 %>% mutate(has_child = ifelse(children > 0, "YES",
"NO"))
```

We then add another dimension to the previous graphical representation i.e whether or not the respondent has a child in the household or not, in an attempt to determine if having children leads to more cases of depressive disorders at various income levels.

```
income_depression_child_barplot = ggplot(Dataframe3, aes(x = has_child,
fill = addepev2)) + geom_bar(position = "fill") + facet_grid(.~income2)
+xlab("income level` and child in household(YES/NO)") + ylab("Proportion
with Depressive Disorder") + scale_fill_brewer(name="Depressive Disorder",
palette = "Paired")
income_depression_child_barplot
```



From the above graph, we can see that at several income levels i.e $15,000 - $50,000 the number of respondents ever diagnosed with depressive disorder is higher in households with children when compared to respondents with no children in the household. There is no statistical difference between respondents with depressive disorder with or without children in the household above $50,000 income level.
This correlation is not conclusive as this trend could simply be due to increase number of dependents in the household.