Exploring the BRFSS data

Setup

Load packages

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

Load data

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

Part 1: Data

On investigation of the BRFSS dataset, first thing observed was that all US states (including District of Couloumbia, Guam, Puerto Rico & US Virgin ISlands) were considered as a sample, of which consists of only adults (18 & older) from the population. These adults were thus, selected Randomly from the sample which means the scope of inference is generalizablility. Causality can be infered only when the data is subjected to Random assigment, but in this case there is no mention of treatment or control of variables (explanatory), and the data is based on blocking principle wherein the data is collected as is. Thus, Causality can be dismissed. The data was collected from a random sample using two methods: Landline phone surveys & Cellular phone surveys. Landline surveys were in turn subjected to further random selection of an adult in the household whereas, cellular phone surveys were further subjected to adults living in a private residence or leasing. Since, the mentioned method of collecting Data via telephone was random selection, it could lead to Non-response bias due to unavailabilty to take calls or Voluntary bias because individuals may choose not to take the survey.

Part 2: Research questions

Research quesion 1:

To find the relation between each state versus the number of adults who were told their cholestrol levels were high, and eventually find out how many of these adults were male or female. Therefore, we concentrate on the following THREE variables 'X_state' , 'toldhi2' and 'sex'. We need to remove the NA values, hence "filter(!is.na())" is used".

```
brfss2013 %>%
 filter(!is.na(toldhi2)) %>%
     filter(!is.na(sex)) %>%
 group_by(X_state, toldhi2, sex) %>%
 summarise (count = n())
```

```
## # A tibble: 212 x 4
## # Groups: X_state, toldhi2 [?]
##    X_state toldhi2 sex    count
##    <fctr>  <fctr>  <fctr> <int>
##  1 Alabama Yes     Male     978
##  2 Alabama Yes     Female  1931
##  3 Alabama No      Male     910
##  4 Alabama No      Female  1813
##  5 Alaska  Yes     Male     717
##  6 Alaska  Yes     Female   775
##  7 Alaska  No      Male     880
##  8 Alaska  No      Female  1244
##  9 Arizona Yes     Male     655
## 10 Arizona Yes     Female   945
## # ... with 202 more rows
```

So, we further filter according to Males who have been told their cholerstrol level was high; in the same way filter according to Females. And thus, using “filter” function; we can compare the relationship between Males vs Females who were told their cholestrol levels were high.

```
brfss2013 %>%
 group_by( X_state, toldhi2, sex ) %>%
    summarise (count = n()) %>%
 filter(sex == "Male" ) %>%
 filter(toldhi2 == "Yes")
```

```
## # A tibble: 53 x 4
## # Groups: X_state, toldhi2 [53]
##    X_state              toldhi2 sex    count
##    <fctr>               <fctr>  <fctr> <int>
##  1 Alabama              Yes     Male     978
##  2 Alaska               Yes     Male     717
##  3 Arizona              Yes     Male     655
##  4 Arkansas             Yes     Male     852
##  5 California           Yes     Male    1691
##  6 Colorado             Yes     Male    2097
##  7 Connecticut          Yes     Male    1230
##  8 Delaware             Yes     Male     810
##  9 District of Columbia Yes     Male     789
## 10 Florida              Yes     Male    5566
## # ... with 43 more rows
```

```
brfss2013 %>%
 group_by(X_state, toldhi2, sex) %>%
     summarise (count = n()) %>%
     filter(sex == "Female" ) %>%
 filter(toldhi2 == "Yes")
```

```
## # A tibble: 53 x 4
## # Groups: X_state, toldhi2 [53]
##    X_state              toldhi2 sex    count
##    <fctr>               <fctr>  <fctr> <int>
##  1 Alabama              Yes     Female  1931
##  2 Alaska               Yes     Female   775
##  3 Arizona              Yes     Female   945
##  4 Arkansas             Yes     Female  1347
##  5 California           Yes     Female  2091
##  6 Colorado             Yes     Female  2567
##  7 Connecticut          Yes     Female  1651
##  8 Delaware             Yes     Female  1267
##  9 District of Columbia Yes     Female  1125
## 10 Florida              Yes     Female  8642
## # ... with 43 more rows
```

Research quesion 2:

To determine the relationship between Income Level and having Any Health Care Coverage. The concerned TWO Variables in this case are “income2” and “hlthpln1”. We use “Filter” function to get rid of NA values.

```
brfss2013 %>%
 filter(!is.na(hlthpln1)) %>%
 filter(!is.na(income2)) %>%
 group_by(income2, hlthpln1) %>%
 summarise(count = n())
```

```
## # A tibble: 16 x 3
## # Groups: income2 [?]
##    income2           hlthpln1  count
##    <fctr>            <fctr>    <int>
##  1 Less than $10,000 Yes       18732
##  2 Less than $10,000 No         6551
##  3 Less than $15,000 Yes       21143
##  4 Less than $15,000 No         5558
##  5 Less than $20,000 Yes       26695
##  6 Less than $20,000 No         8061
##  7 Less than $25,000 Yes       33312
##  8 Less than $25,000 No         8295
##  9 Less than $35,000 Yes       41738
## 10 Less than $35,000 No         7024
## 11 Less than $50,000 Yes       55575
## 12 Less than $50,000 No         5824
## 13 Less than $75,000 Yes       61732
## 14 Less than $75,000 No         3414
## 15 $75,000 or more   Yes      113023
## 16 $75,000 or more   No         2771
```

Also, we can seperately find the adults having/not having any Health Plan with respect to their income levels, as follows:

```
brfss2013 %>%
 filter(!is.na(hlthpln1)) %>%
 filter(!is.na(income2)) %>%
 group_by(income2, hlthpln1) %>%
 summarise(count = n()) %>%
 filter(hlthpln1 == "Yes")
```

```
## # A tibble: 8 x 3
## # Groups: income2 [8]
##   income2           hlthpln1  count
##   <fctr>            <fctr>    <int>
## 1 Less than $10,000 Yes       18732
## 2 Less than $15,000 Yes       21143
## 3 Less than $20,000 Yes       26695
## 4 Less than $25,000 Yes       33312
## 5 Less than $35,000 Yes       41738
## 6 Less than $50,000 Yes       55575
## 7 Less than $75,000 Yes       61732
## 8 $75,000 or more   Yes      113023
```

```
brfss2013 %>%
 filter(!is.na(hlthpln1)) %>%
 filter(!is.na(income2)) %>%
 group_by(income2, hlthpln1) %>%
 summarise(count = n()) %>%
 filter(hlthpln1 == "No")
```

```
## # A tibble: 8 x 3
## # Groups: income2 [8]
##   income2           hlthpln1 count
##   <fctr>            <fctr>   <int>
## 1 Less than $10,000 No        6551
## 2 Less than $15,000 No        5558
## 3 Less than $20,000 No        8061
## 4 Less than $25,000 No        8295
## 5 Less than $35,000 No        7024
## 6 Less than $50,000 No        5824
## 7 Less than $75,000 No        3414
## 8 $75,000 or more   No        2771
```

Research quesion 3:

To find out how many Single(Never Married) adults own or rent a home and determine the relationship in accordance to the Number of hours per day they work. Thus, we concentrate of the following THREE Variables: “Marital”, “renthom1” and “scntwrk1”.

First, we find the relationship between marital status & own/rent a home, with respect to Number of hours per day they work using “filter(!is.na())” command to filter out the NA values.

```
brfss2013 %>%
 filter(!is.na(marital)) %>%
 filter(!is.na(renthom1)) %>%
 filter(!is.na(scntwrk1)) %>%
 group_by(marital, renthom1, scntwrk1) %>%
summarise(count = n())
```

```
## # A tibble: 912 x 4
## # Groups: marital, renthom1 [?]
##    marital renthom1 scntwrk1 count
##    <fctr>  <fctr>      <int> <int>
##  1 Married Own             0     2
##  2 Married Own             1     7
##  3 Married Own             2    17
##  4 Married Own             3    18
##  5 Married Own             4    34
##  6 Married Own             5    50
##  7 Married Own             6    25
##  8 Married Own             7    15
##  9 Married Own             8    68
## 10 Married Own             9    14
## # ... with 902 more rows
```

Now, we can further find out, specific data like: How many “Never married” adults, “own a home” given that they work “45 or less hours per day”? Following is the code to run:

```
brfss2013 %>%
 filter(!is.na(renthom1)) %>%
 filter(!is.na(scntwrk1)) %>%
 filter(!is.na(marital)) %>%
 group_by(marital, renthom1, scntwrk1) %>%
 filter(marital == "Never married") %>%
 filter(renthom1 == "Own") %>%
 filter(scntwrk1 <= 45) %>%
summarise(count = n())
```

```
## # A tibble: 45 x 4
## # Groups: marital, renthom1 [?]
##    marital       renthom1 scntwrk1 count
##    <fctr>        <fctr>      <int> <int>
##  1 Never married Own             0     2
##  2 Never married Own             1     1
##  3 Never married Own             2     2
##  4 Never married Own             3     4
##  5 Never married Own             4     8
##  6 Never married Own             5     2
##  7 Never married Own             6     6
##  8 Never married Own             7     3
##  9 Never married Own             8    13
## 10 Never married Own             9     2
## # ... with 35 more rows
```

Similarly, we can find out the “Never married” adults, “rent a home”, just by changing the filter: “filter(renthom1 ==”Rent“)”

Part 3: Exploratory data analysis

Research quesion 1:

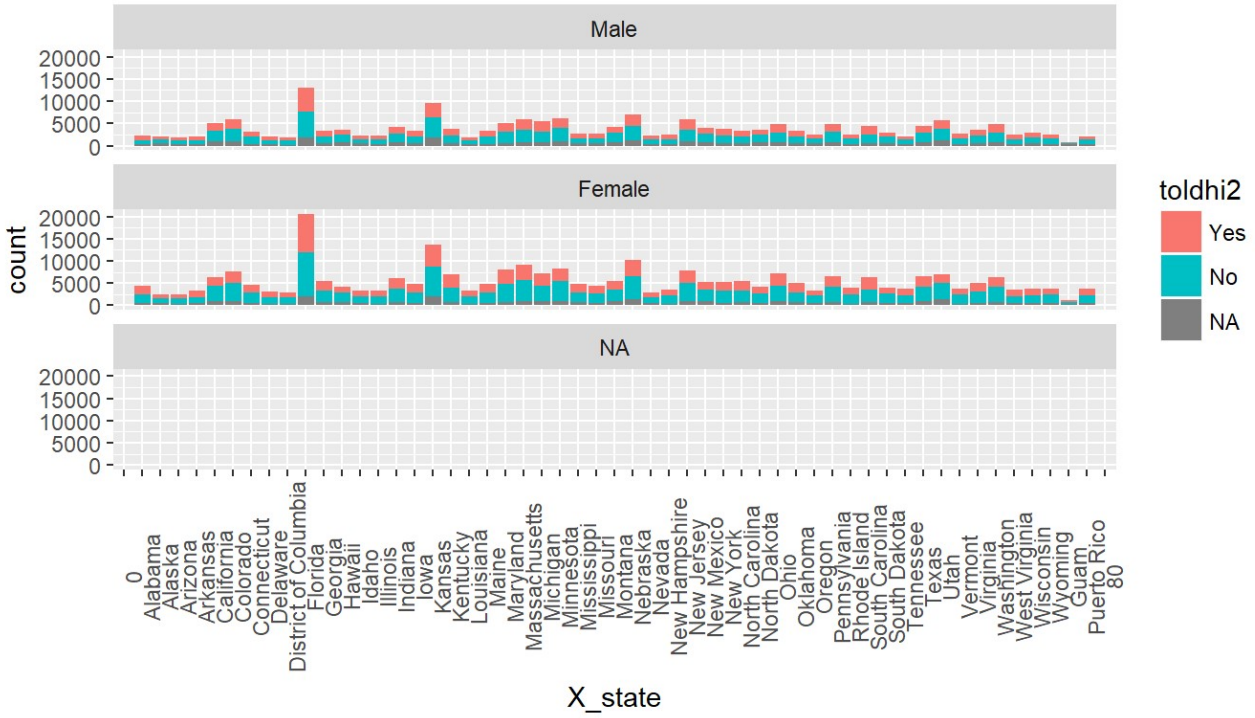
According to this question, there are #3 CATEGORICAL variables involved ie., “X_state”, “toldhi2” & “sex”. For EDA in this case, we can use BAR PLOT with FACETS, and we achieve that using the below code. We can see below, it is a diverse output considering the Number of states. P.S.: Refer the research questions for numeric analysis & table.

Firstly, there is this generalized observation that the bar-plot output resembles a right-skewed shape, wherein the “State of FLORIDA” has the highest amount of adults who were told their “cholestrol levels are HIGH”; subsequently if we COMPARE MALE vs FEMALE, thiese Levels of cholestrol is even HIGHER. At the same we can observe that there a considerable amount of people who have been told their cholestrol levels are NOT HIGH in the same state as well!

The second most highest cholestrol levels are observed in the State of KANSAS, where there are more number of FEMALES who DID NOT have HIGH levels of cholestrol. Whereas, the LEAST number of males & females who were told their cholestrol levels were high were in the “SATE OF GUAM”.

From this EDA, Overall we can observe that, MORE number of FEMALES have been told they had HIGH cholestrol levels than MALES in the respective STATES.

```
ggplot(data = brfss2013, aes(x=X_state, fill= toldhi2) ) +
 geom_bar() + facet_wrap( ~ sex, ncol= 1) + theme(axis.text.x = element_text(angle = 9
0))
```



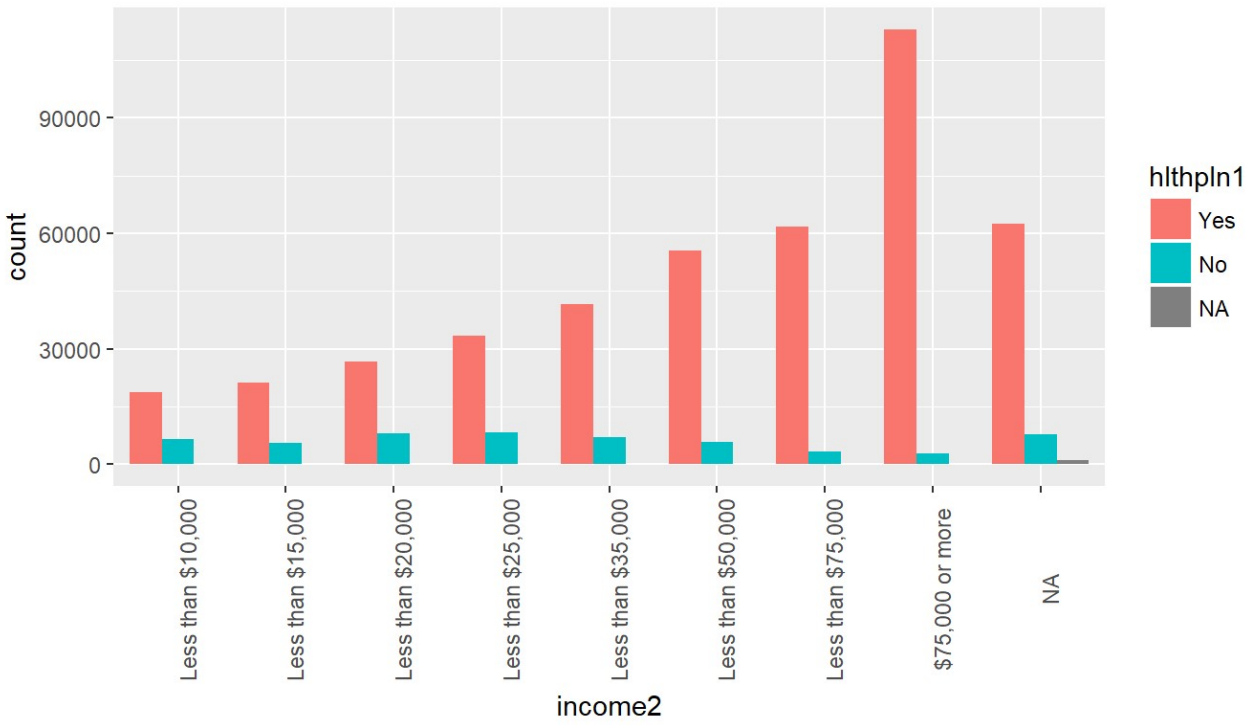
Research quesion 2:

In this question we are dealing with TWO CATEGORICAL variables ie., “income2” and “hlthpln1”. Thus, we can find the relationship between these TWO using BAR-PLOT using the below commands. Looking at the Bar-plot we can read the results better if we use position “dodge” for the geom_bar command. P.S.: Refer the research questions for numeric analysis & table.

So, we observe that the BAR_PLOT is LEFT-SKEWED in SHAPE (disregarding the NA values), wherein the adults having an income “$75,000 or MORE”" MOST DEFINITELY HAVE ANY HEALTH CARE COVERAGE. Whereas, those having INCOME “Less than $10,000”, have LESSER access to any HEALTH CARE COVERAGE. At the same time, the trends for those adults having NO HEALTH CARE COVERAGE vs INCOME show a RIGHT-SKEWED shape (disregaring the NA values).

Thus, the EDA suggests that, those adults having HIGHER INCOME have better chances of HAVING any HEALTH CARE COVERAGE.

```
ggplot(brfss2013, aes(x = income2 , fill = hlthpln1)) +
  geom_bar(position = "dodge") + theme(axis.text.x = element_text(angle = 90))
```



Research quesion 3:

This question is concerned with “2 CATEGORICAL & 1 NUMERICAL VAriables” i.e., “marital” and “renthom1” as Categorical & “scntwrk1” as Numerical variables. P.S.: Refer the research questions for numeric analysis & table.

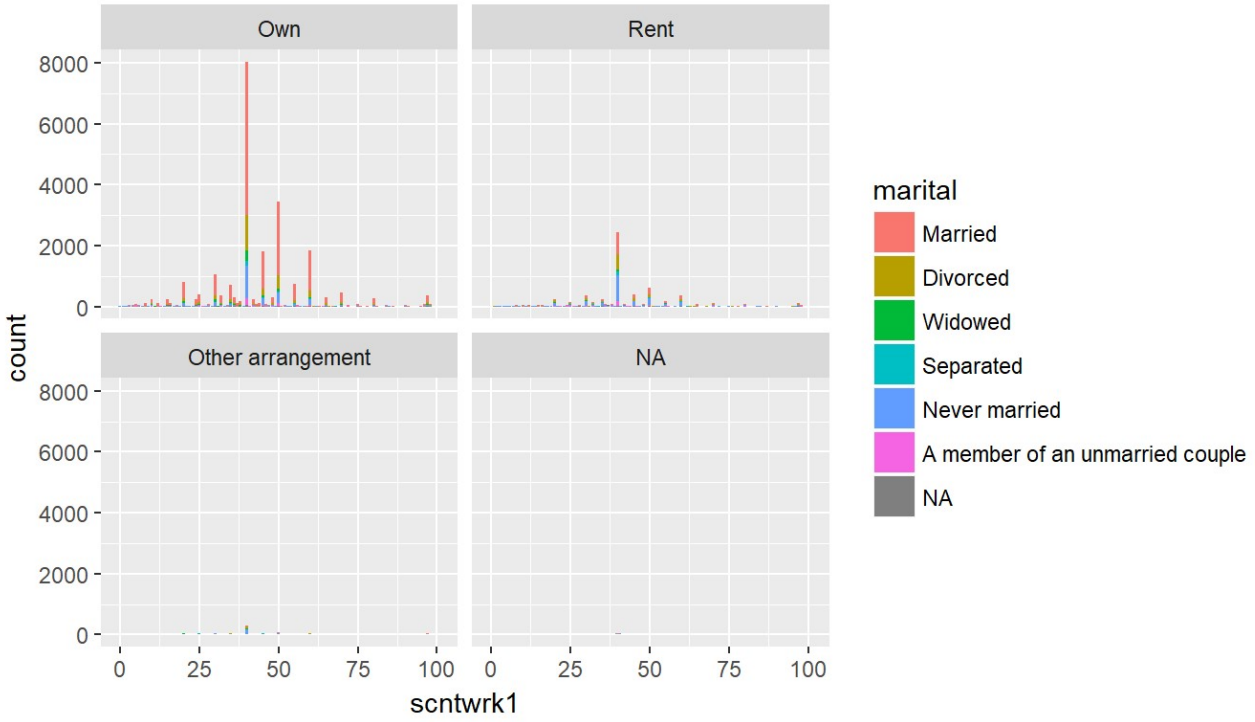
Thus, we can find the relationship between these THREE using “BAR-PLOT with FACETS” using the below commands. Looking at the bar-plot in general, we are unable to see the plots clearly due the VAST range of Variable values for “Number of hrs/week you worked”(scntwrk1). Thus, we can observe that, there is a “UNIMODAL” but, SOMEWHAT “RIGHT-SKEWED” shape to the Bar-plot suggesting that :- 1. MORE “MARRIED” adults, who work for “ABOUT 40 HRS per WEEK” are the ones who “OWN a HOME”, and the category of “A MEMBER OF A MARRIED COUPLE” working about 40 HRS have LEAST amount of people OWNING a HOME. 2. COMPARE that to the adults who “RENT A HOME”, again we see similar trends in which, MORE “MARRIED” adults, who work for “ABOUT 40 HRS per WEEK” show the TOP results. 3. Simiar is the case with adults having “Other arrangement” for home, but the count of ADULTS in this CATEGORY is comparatively LESS, hence we can hardly see the PLOTS.

```
ggplot(brfss2013, aes(x=scntwrk1, fill = marital)) +
  geom_bar() + facet_wrap( ~ renthom1, ncol= 2)
```

```
## Warning: Removed 459413 rows containing non-finite values (stat_count).
```

```
## Warning: position_stack requires non-overlapping x intervals

## Warning: position_stack requires non-overlapping x intervals
```



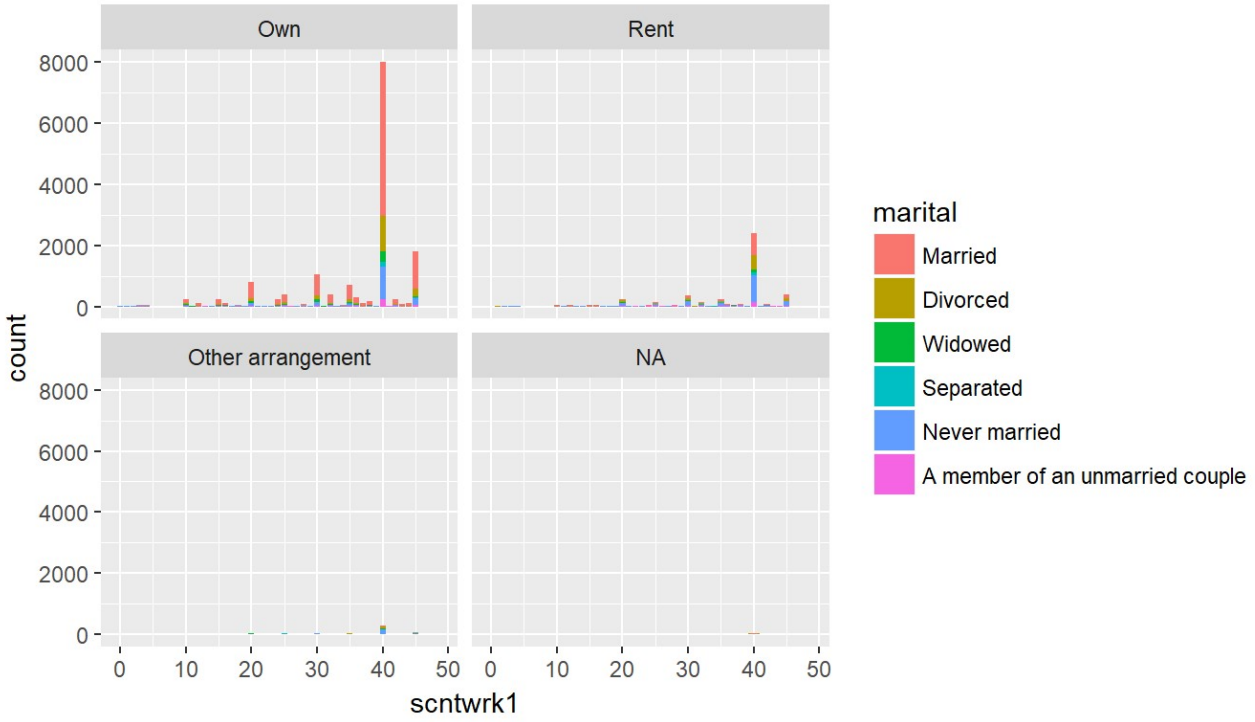
TO have a clear understanding of the BAR-Plot we can divide the observations in a couple of parts. Here, I have considered “Number of hrs per WEEK worked” to be UPTO 45 or LESS, having the following RESULTS.

As observed before, we can support the results from theprevious observations that, the PLOT is “RIGHT-SKEWED” in case of “OWNING a HOME”, “RENTING a HOME” & “Having OTHER ARRANGEMENT”. Also, MORE “MARRIED” adults “working for 45 hours or LESS”, “OWN a HOME” or “RENT”.

Overall,from this EDA we can conclude that MARRIED adults have a higher possibility of OWNING or RENTING a HOME.

```
brfss2013_marital_rent_hrspwk <- brfss2013 %>%
 filter(!is.na(marital), scntwrk1 <="45") %>%
 select(marital, scntwrk1, renthom1)
 brfss2013_marital_rent_hrspwk %>%
ggplot( aes(x=scntwrk1, fill = marital)) +
  geom_bar() + facet_wrap( ~ renthom1, ncol= 2)
```

```
## Warning: position_stack requires non-overlapping x intervals
```


Own
Rent
Other arrangement
NA
count
scntwrk1
0
2000
4000
6000
8000
0
10
20
30
40
50
marital
Married
Divorced
Widowed
Separated
Never married
A member of an unmarried couple