Exploring the BRFSS data

Intzar Singh

Setup

Load packages

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

Load data

```
load("/Users/Intzar/Downloads/brffs2013.rd")
```

Part 1: Data

The data collected in this survey used a random sampling technique. This allows us to see that the results and valid interpretations of this data can be generalized to a larger population. However, we need to note that these results should only be generalized to the populations that this sample is representative of.

Furthermore, we can also rule out any causal factors in this study because there is no control group to compare to. In addition to this, observational studies, which this survey can be categorized as, can not determine causal relationships.

However, we may be able to get information in this data that may help us determine weather certain factors deserved to be looked at in more detail. These topics of further study could greatly aide the CDC in its efforts to predict upcoming health trends for communities.

For more Information on the data collection process as provided by the CDC: https://www.cdc.gov/brfss/annual_data/2019/pdf/overview-2019-508.pdf

---

Part 2: Research questions

Research quesion 1: The first question that is of interest is to determine whether there is any relationship between the marital status of a respondent, the number of children in their home, and the number of days in a month that they felt they were bad mentally.

This question could help us determine if respondents that were married with children felt more positive about their mental health when compared to married respondents without children. On the opposite end of

this question, we will also be able to see if any relationships may exist between unmarried respondents with kids, and their mental health.

More formally :

Do respondents with children feel differently about their mental health based on their Marital Status? Furthermore, can we determine a trend for married and unmarried respondents mental health and the number of children they have?

Research question 2: Next I want to evaluate the relationship between education level and likelihood of owning a home. To do this, we will use income to categorize our respondents and isolate the relationship to be between education and home ownership.I believe that income would definitely be a factor in home ownership rates, but that question is not of interest.

More formally:

When income brackets are equalized, are college graduates more likely to own homes than non college graduates? We can also do some minor analysis along the way to see the discrepancy in income between non college graduates and college graduates.

Research question 3: The next research question we would like to explore is the amount of times someone has gone to the doctor in the past year, based on whether they wear their seat belt and have gotten their flu shot.

What I am interested in testing here, is a parallel in being responsible about your health and safety across different fields. Although we may get things that are nonsensical or completely unrelated, I am interested in seeing whether someone who wears their seat belt and gets their flu shot, is also more likely to visit the doctor.

More formally:

Do respondents that wear their seat belt and have gotten a flu shot visit the doctor more than respondents that do not fit that criteria? * * *

Part 3: Exploratory data analysis

Research question 1: We are going to first filter the data we need out of the bigger set of data, and clean it up to get rid of any NA values if they exist.

```
q1_data <- brfss2013 %>%
  select(marital, children, menthlth)

q1_data <- q1_data %>%
  filter_at(vars(marital,children,menthlth), all_vars(!is.na(.)))
```

This is all the data that we really want to clean, so we can now move over into looking at some statistics about the data, based on wether the participant is married or not.

```
q1_data %>%
  group_by(marital) %>%
  summarise(ment_med = median(menthlth), ment_mean= mean(menthlth))
```

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

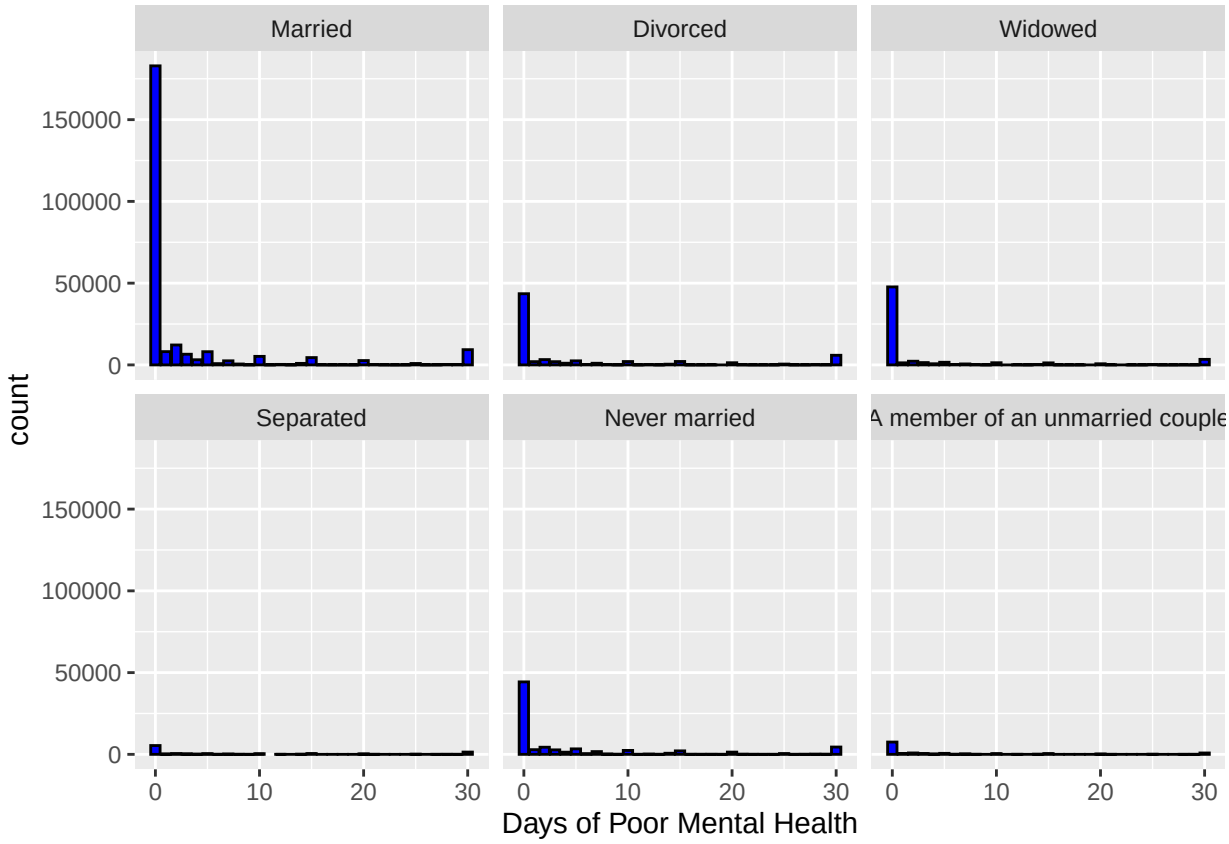
```
## # A tibble: 6 x 3
##   marital                        ment_med ment_mean
```

```
##   <fct>                                <dbl>     <dbl>
## 1 Married                                  0      2.59
## 2 Divorced                                 0      4.81
## 3 Widowed                                  0      3.04
## 4 Separated                                0      7.19
## 5 Never married                            0      4.25
## 6 A member of an unmarried couple          0      4.49
```

From this table we can see that although the data for all members of the marital group has a median of 0, their means have quite a bit of variation between them. We can see that people who are married have a lower average number of days they felt poorly about their mental health for their group. Furthermore, we can see that separated individuals post the highest average.

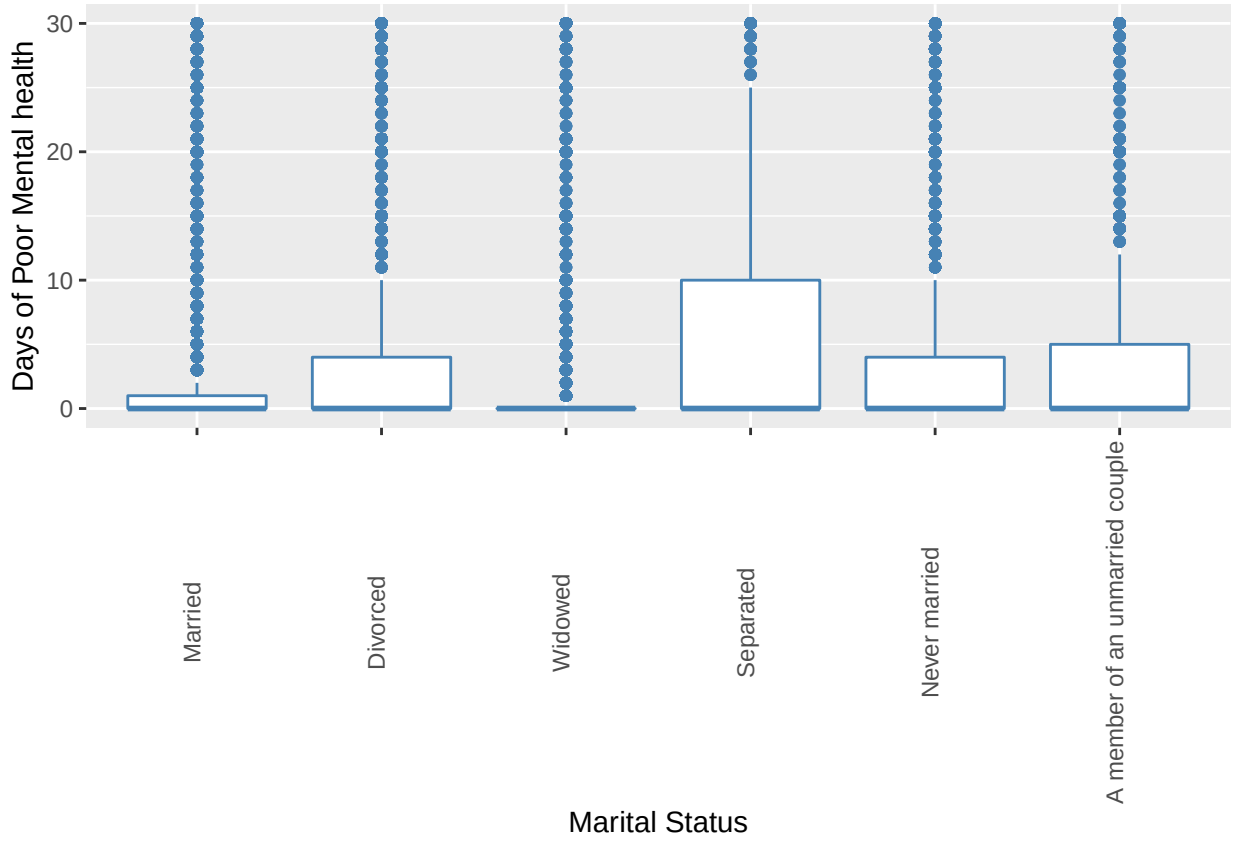
I now want to look at the distribution of the mental health variable for each of these groups and see if we can find something that is of interest to us.

```
ggplot(data = q1_data) +
    geom_bar(aes(x = menthlth), colour = "black", fill = "blue") +
    xlab("Days of Poor Mental Health")+
    facet_wrap(~ marital)
```



We can see from the above frequency plot that all of the various distributions for each type of data are right skewed. Furthermore, the median of all our distributions seems to be 0, from the graph as well, which can be verified by looking at the summary table above.

```
ggplot(data = q1_data, aes(x= marital, y= menthlth)) +
  xlab("Marital Status")+
  ylab("Days of Poor Mental health")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))+
  geom_boxplot(color = "steelblue")
```



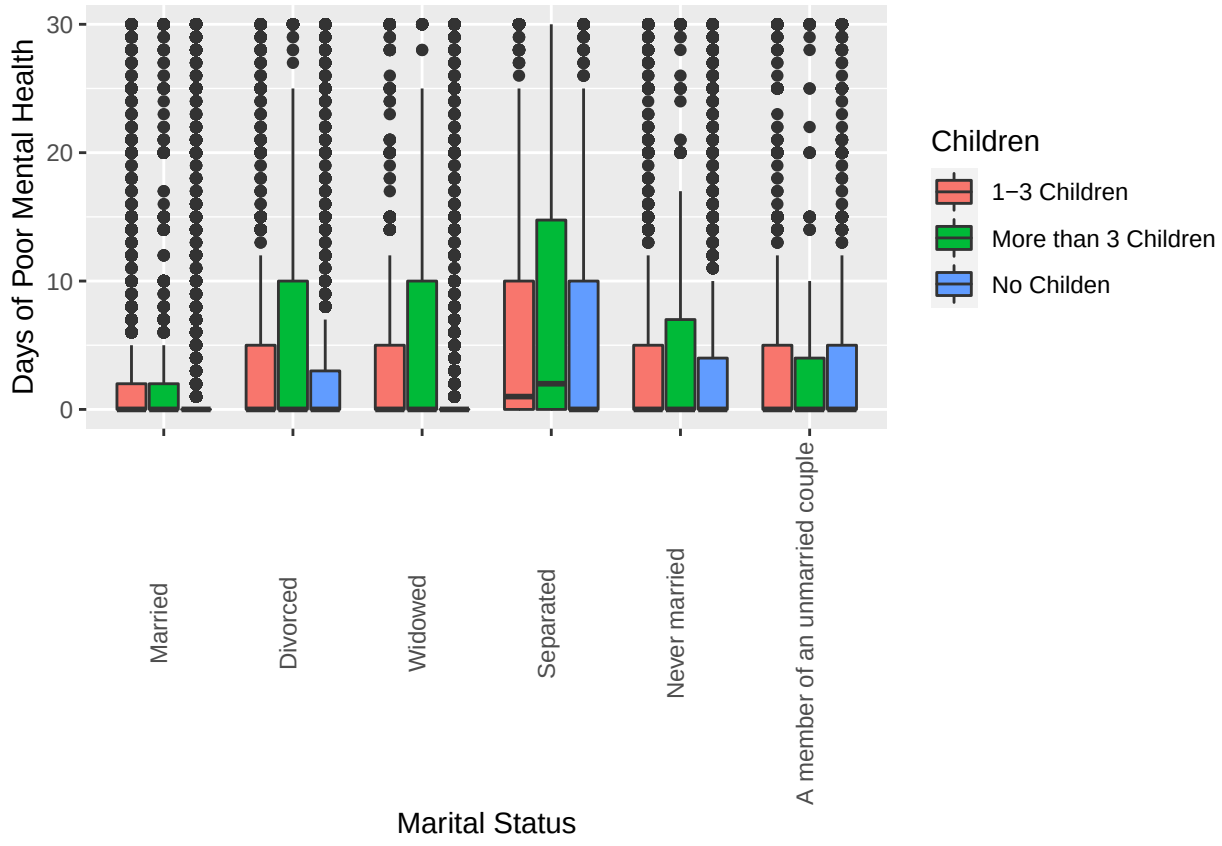
Insights:From the Frequency Plot, and the Box Plot we can see that some variables might require some further studying. One case of this can be seen in the Separated and Divorced cases. The wide range of the mental health variable for Separated couples when compared to those that are divorced, or married, may suggest that there are some factors that may be leading to more stress. Furthermore, When we compare the means of the data, we can further see that there is definitely some discrepancy between Divorced participants and their feelings on their mental Health, and those of Married Participants. One Key thing to note however, is that the median for all of these categorical variables was the same. What this suggests to me is that those participants who are Separated or Divorced and express poor mental health for at least 1 day of the month, will express a greater number on average for participants of the same type who are married.

We can now turn over into evaluating how the number of children that a participant may have affects the number of days they express poor mental health. Ideally, we will get a working version of a visualization that will allow us to establish some type of relationship between these variables.

```
q1_data <- q1_data %>%
  mutate(cat_child = ifelse(children<1,"No Childen",ifelse(children<=3, "1-3 Children","More than 3 Chil
```

```
ggplot(data = q1_data, aes( x= marital, y= menthlth, fill = cat_child))+
  geom_boxplot()+
  xlab("Marital Status")+
```

```
  ylab("Days of Poor Mental Health")+
  guides(fill = guide_legend("Children"))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))
```



We can see from this graph that when we look at the Marital Status categories that we have decided are of importance to us, we can see that Seperated participants that had children had a higher median for the amount of days that they felt poor mental health. Furthermore, Separated participants with more than 3 children had not only the highest median, but also the largest middle 50th percentile.

```
q1_data %>%
  group_by(cat_child,marital) %>%
  summarise(ment_med = median(menthlth), ment_mean= mean(menthlth), count = n())
```

```
## `summarise()` regrouping output by 'cat_child' (override with `.groups` argument)
```

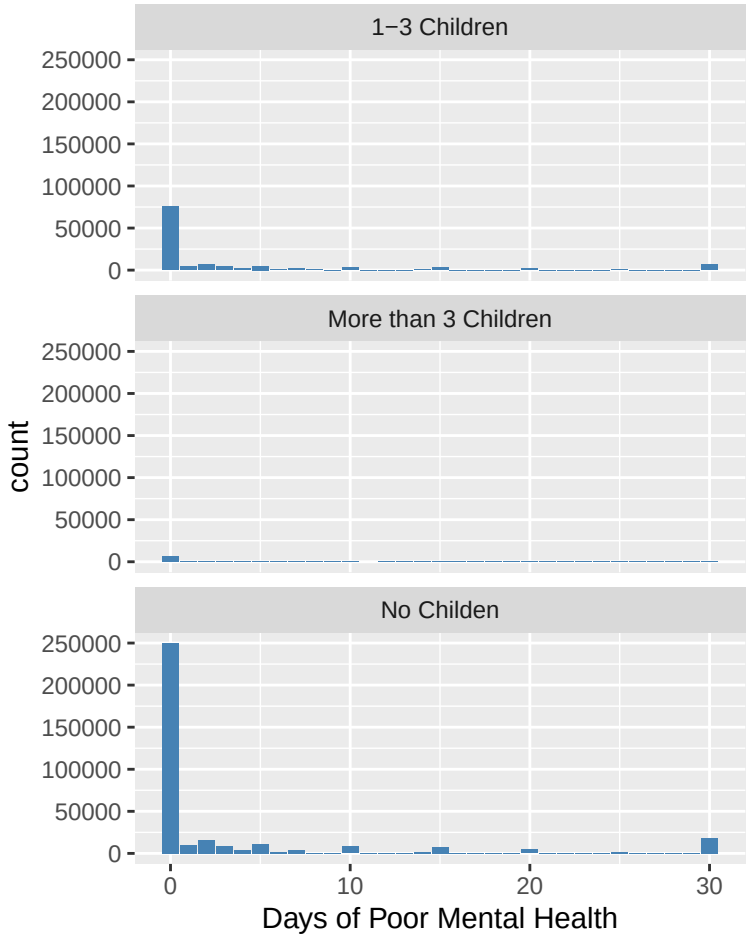
```
## # A tibble: 18 x 5
## # Groups:   cat_child [3]
##    cat_child           marital                        ment_med ment_mean  count
##    <chr>               <fct>                             <dbl>     <dbl>  <int>
##  1 1-3 Children        Married                               0      2.87  76681
##  2 1-3 Children        Divorced                              0      5.56  13139
##  3 1-3 Children        Widowed                               0      5.17   3234
##  4 1-3 Children        Separated                             1      7.25   3695
##  5 1-3 Children        Never married                         0      4.84  16479
##  6 1-3 Children        A member of an unmarried coup~        0      4.70   4707
```

```
##  7 More than 3 Children Married                                  0      3.06   7096
##  8 More than 3 Children Divorced                                 0      6.60    762
##  9 More than 3 Children Widowed                                  0      6.20    201
## 10 More than 3 Children Separated                                2      7.69    398
## 11 More than 3 Children Never married                            0      5.44   1245
## 12 More than 3 Children A member of an unmarried coup~           0      4.42    431
## 13 No Childen           Married                                  0      2.43 165804
## 14 No Childen           Divorced                                 0      4.60  54883
## 15 No Childen           Widowed                                  0      2.92  60055
## 16 No Childen           Separated                                0      7.13   6265
## 17 No Childen           Never married                            0      4.05  55887
## 18 No Childen           A member of an unmarried coup~           0      4.37   7851
```

This table further shows us the discrepancies in how children affect the mean and median mental health variable for various individuals.

Lastly I want to see the distribution of the Mental Health Variable grouped by how many children a person has. This will allow me to see whether we have a substantial amount of respondents to make a judgment on if these factors require further studying.

```
ggplot(q1_data)+
  geom_bar(aes(x=menthlth), fill = "steelblue")+
  facet_wrap( ~ cat_child, ncol=1)+
  xlab("Days of Poor Mental Health")
```



From the graph above, we can see that the distributions themselves look very similar, the only major difference between them being the total number of respondents that fit into these categories. Because these distributions are similar, I believe that as we explored earlier, the marital status of a person is having an effect on the respondents mental health.

As a whole, I do believe that looking at the data in this perspective has helped us look into how the different Marital Status variables interact with the other two variables of our analysis. If I were to take this past the EDA phase, I would focus in on the Separated, Married, and Divorced Categories and see if there can be any inferences drawn based on further study.

Research question 2: We are going to first filter the data we need out of the bigger set of data, and clean it up to get rid of any NA values if they exist.

When income brackets are taken into consideration, are college graduates more likely to own homes than non college graduates? We can also do some minor analysis along the way to see the discrepancy in income between non college graduates and college graduates.

```
q2_data <- brfss2013 %>%
  select(X_incomg,renthom1,X_educag) %>%
  filter_at(vars(X_incomg,renthom1,X_educag),all_vars(!is.na(.)))

q2_data %>%
  group_by(X_incomg) %>%
  summarise( count = n())
```

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

```
## # A tibble: 5 x 2
##   X_incomg                          count
##   <fct>                             <int>
## 1 Less than $15,000                 51602
## 2 $15,000 to less than $25,000      75683
## 3 $25,000 to less than $35,000      48412
## 4 $35,000 to less than $50,000      61029
## 5 $50,000 or more                  179887
```

```
q2_data %>%
  group_by(X_educag) %>%
  summarise( count = n())
```

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

```
## # A tibble: 4 x 2
##   X_educag                                    count
##   <fct>                                       <int>
## 1 Did not graduate high school                32873
## 2 Graduated high school                      117645
## 3 Attended college or technical school       115199
## 4 Graduated from college or technical school 150896
```

```
q2_data %>%
  group_by(renthom1) %>%
  summarise( count = n())
```

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

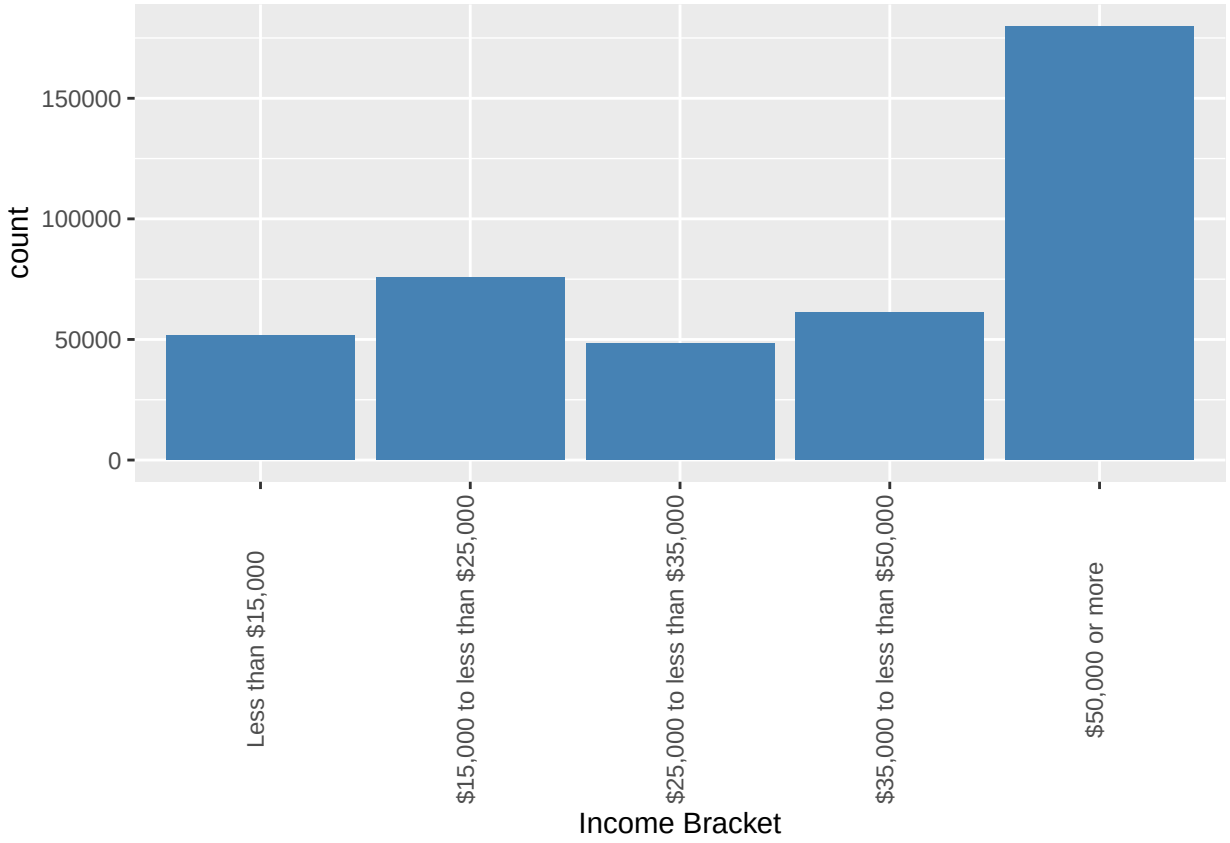
```
## # A tibble: 3 x 2
##   renthom1           count
##   <fct>              <int>
## 1 Own               303769
## 2 Rent               97088
## 3 Other arrangement  15756
```

The above table summaries give us a good idea of the way the variables are laid out across the data set. To get a better visualization of this, we will now create a bar graph.

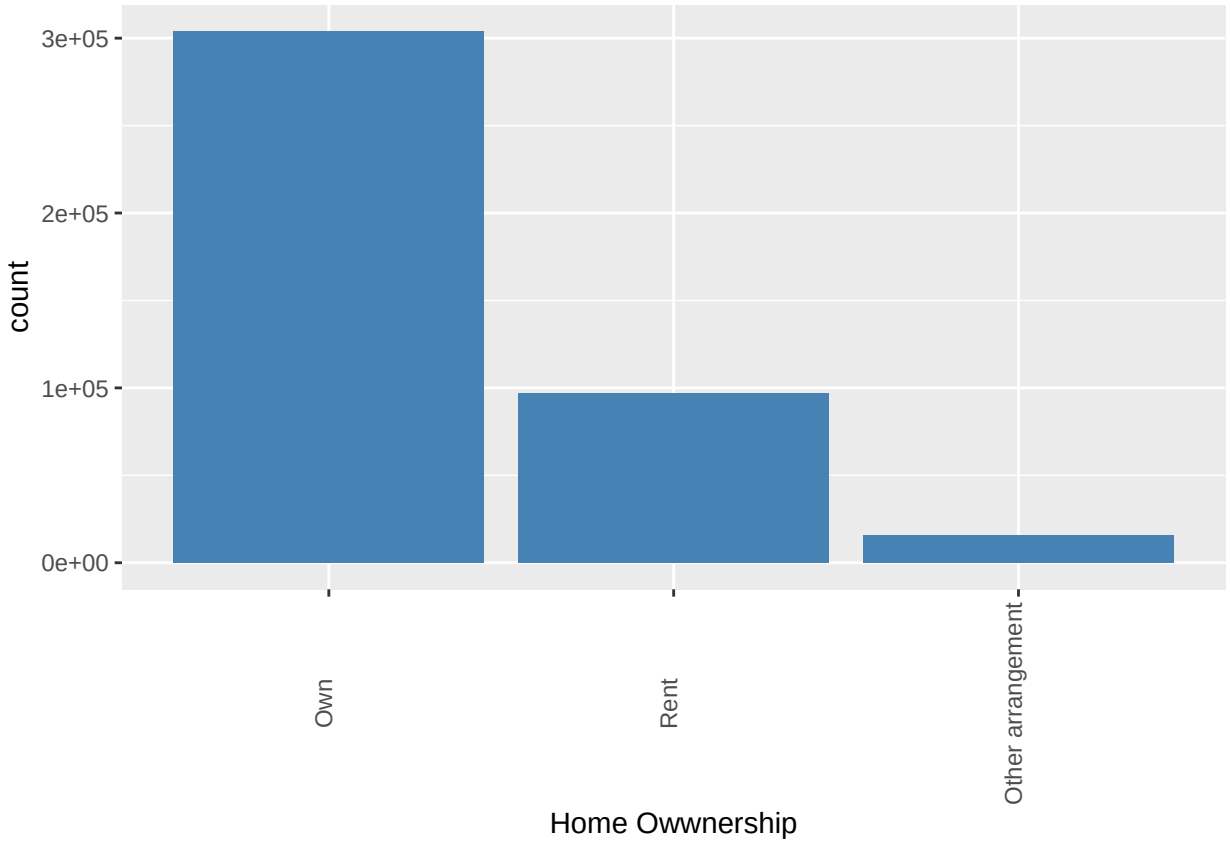
There are some things to note about this question that make our analysis different: - Categorical Data was already provided in the form we need, so no code was needed - All 3 variables being compared are categorical - We cant do our normal Numeric operations and Descriptions because these data tables are all categorical.

Now I want to first plot the distributions of all of these variables. This will be similar to the data tables above, but more visually appealing.

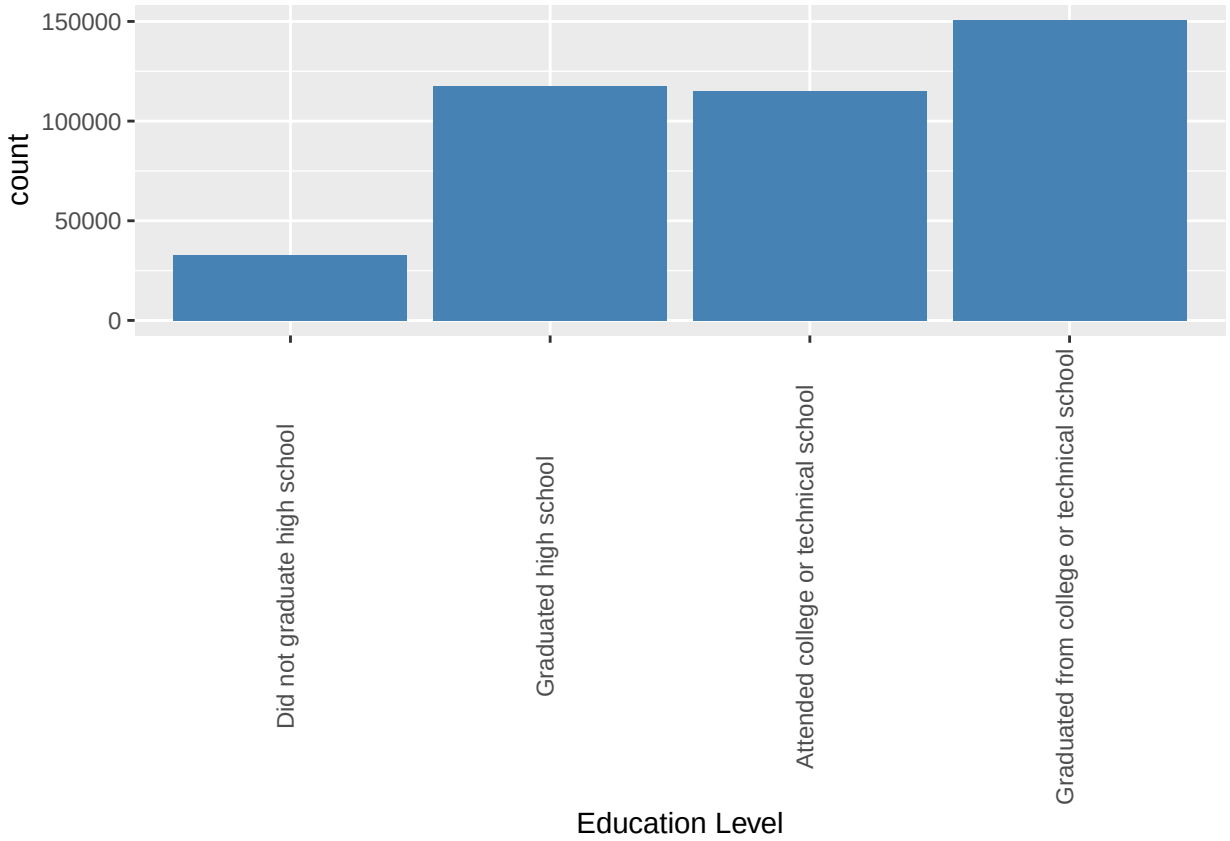
```
ggplot(data = q2_data, aes(x= X_incomg))+
  geom_bar(fill = "steelblue")+
  xlab("Income Bracket")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))
```



```
ggplot(data = q2_data, aes(x= renthom1))+
  geom_bar(fill = "steelblue")+
  xlab("Home Owwnership")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))
```




```
ggplot(data = q2_data, aes(x= X_educag))+
  geom_bar(fill = "steelblue")+
  xlab("Education Level")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))
```



The above bar graphs give us a better visual understanding of the dstribution of our data, but we are still missing some comparative visualizations that can help us with our question.

Since we are dealing with only categorical variables, I will start by showing a contingency table. We will then evaluate the distrubution based on these contigencies with a layered bar graph.

```
q2_data %>%
  group_by(X_educag, X_incomg) %>%
  summarise(count = n())
```

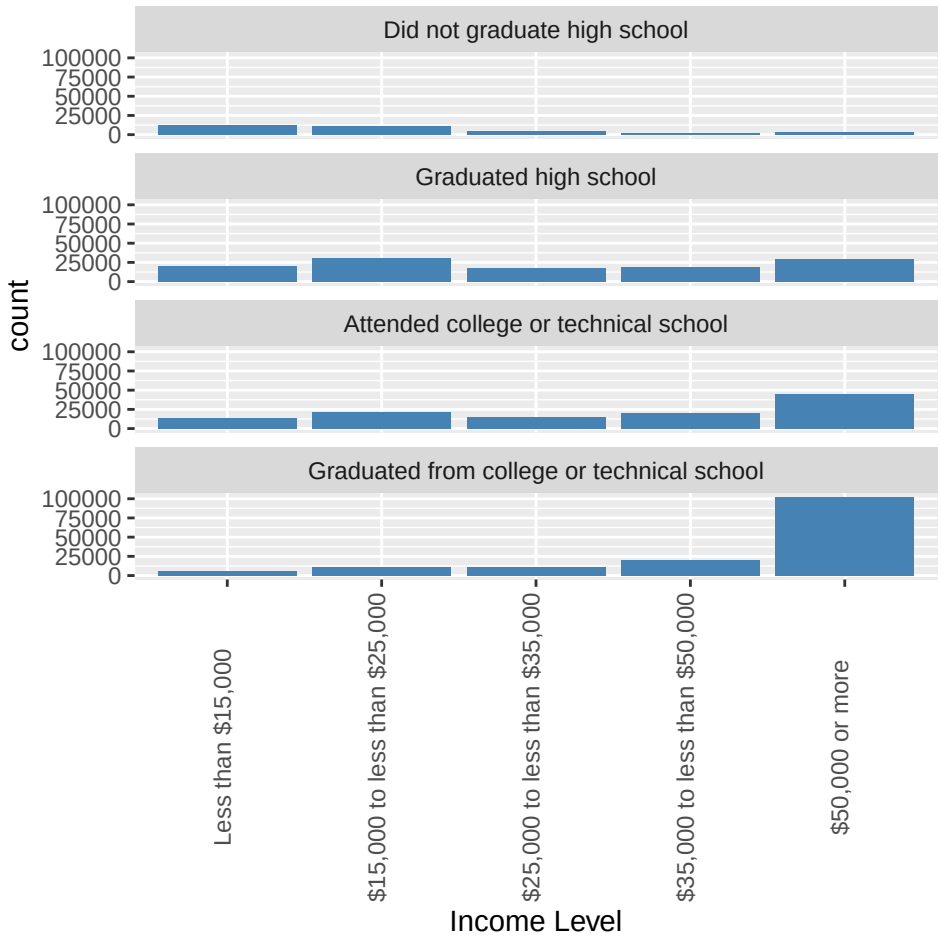
```
## `summarise()` regrouping output by 'X_educag' (override with `.groups` argument)
```

```
## # A tibble: 20 x 3
## # Groups:   X_educag [4]
##    X_educag                                X_incomg                    count
##    <fct>                                   <fct>                       <int>
##  1 Did not graduate high school            Less than $15,000           12641
##  2 Did not graduate high school            $15,000 to less than $25,0~ 11073
##  3 Did not graduate high school            $25,000 to less than $35,0~  4007
##  4 Did not graduate high school            $35,000 to less than $50,0~  2529
##  5 Did not graduate high school            $50,000 or more              2623
##  6 Graduated high school                   Less than $15,000           20045
##  7 Graduated high school                   $15,000 to less than $25,0~ 31190
##  8 Graduated high school                   $25,000 to less than $35,0~ 18003
##  9 Graduated high school                   $35,000 to less than $50,0~ 18722
## 10 Graduated high school                   $50,000 or more             29685
```

```
## 11 Attended college or technical school       Less than $15,000            13068
## 12 Attended college or technical school       $15,000 to less than $25,0~  21922
## 13 Attended college or technical school       $25,000 to less than $35,0~  15298
## 14 Attended college or technical school       $35,000 to less than $50,0~  19671
## 15 Attended college or technical school       $50,000 or more              45240
## 16 Graduated from college or technical school Less than $15,000             5848
## 17 Graduated from college or technical school $15,000 to less than $25,0~  11498
## 18 Graduated from college or technical school $25,000 to less than $35,0~  11104
## 19 Graduated from college or technical school $35,000 to less than $50,0~  20107
## 20 Graduated from college or technical school $50,000 or more             102339
```

The above data is rather messy, the key question we are trying to see, is based on education level, is the proportion of total respondents income higher with more education. We can try a bar graph next.

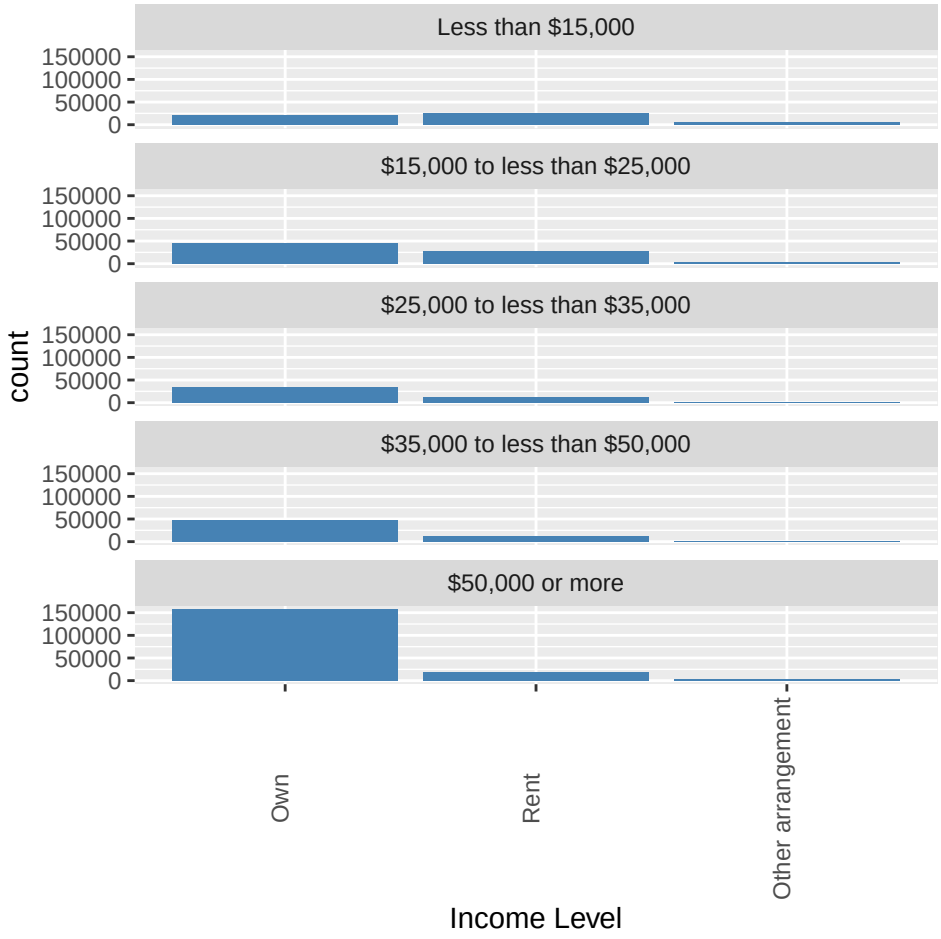
```
ggplot(data=q2_data)+
  geom_bar(aes(x= X_incomg), fill ="steelblue")+
  facet_wrap(~ X_educag, ncol = 1)+
  xlab("Income Level")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))
```



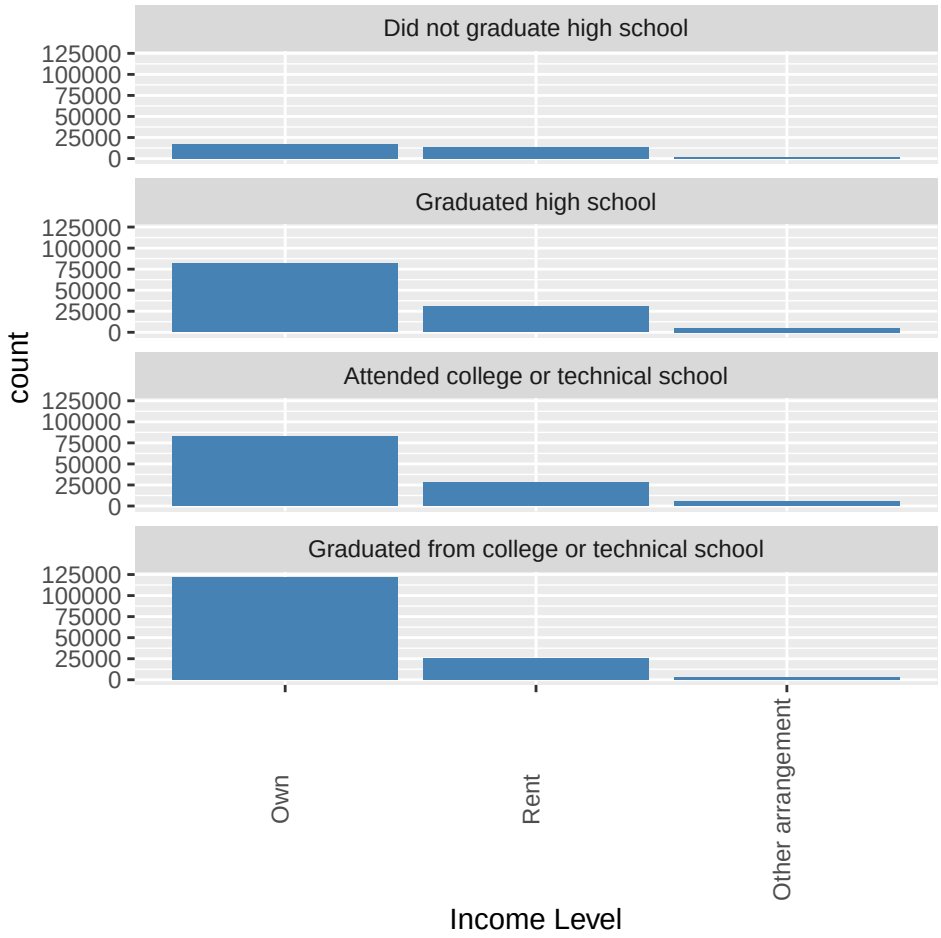
From the above data, it becomes quite clear that as you increase your education level the more likely you are to fall into a higher income category.

Finally, lets look at the way homeownership is related to these two variables, and then lets determine if there is a way to see all three of these variables together.

```
ggplot(data=q2_data)+
  geom_bar(aes(x= renthom1), fill ="steelblue")+
  facet_wrap(~ X_incomg, ncol =1)+
  xlab("Income Level")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))
```




```
ggplot(data=q2_data)+
  geom_bar(aes(x= renthom1), fill ="steelblue")+
  facet_wrap(~ X_educag, ncol = 1)+
  xlab("Income Level")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))
```



In data table above graphs would look like the following, click through to see the various differences

```
q2_data %>%
  group_by(X_educag,renthom1) %>%
  summarise(count = n())
```

```
## `summarise()` regrouping output by 'X_educag' (override with `.groups` argument)
```

```
## # A tibble: 12 x 3
## # Groups:   X_educag [4]
##    X_educag                                renthom1           count
##    <fct>                                   <fct>              <int>
##  1 Did not graduate high school            Own                17446
##  2 Did not graduate high school            Rent               13256
##  3 Did not graduate high school            Other arrangement   2171
##  4 Graduated high school                   Own                81787
##  5 Graduated high school                   Rent               30599
##  6 Graduated high school                   Other arrangement   5259
##  7 Attended college or technical school    Own                82270
##  8 Attended college or technical school    Rent               28074
##  9 Attended college or technical school    Other arrangement   4855
## 10 Graduated from college or technical school Own              122266
## 11 Graduated from college or technical school Rent              25159
```

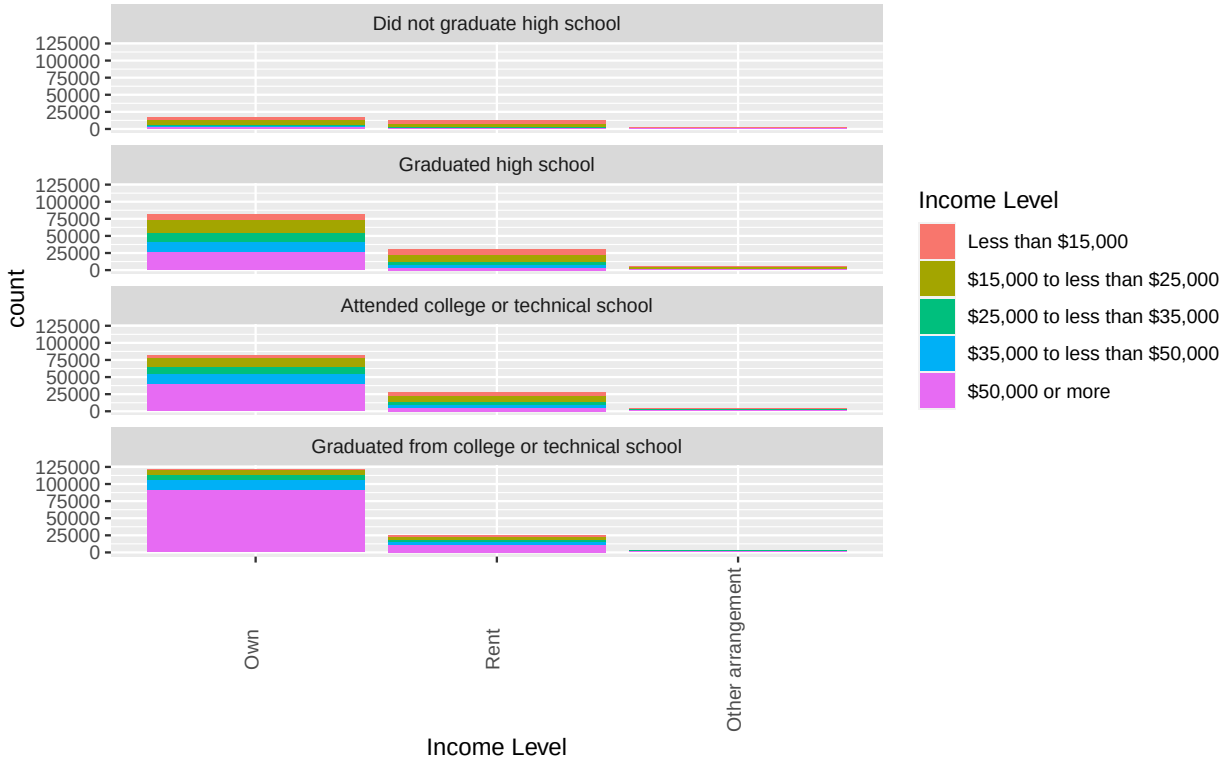
```
## 12 Graduated from college or technical school Other arrangement   3471
```

```
q2_data %>%
  group_by(X_incomg,renthom1) %>%
  summarise(count = n())
```

```
## `summarise()` regrouping output by 'X_incomg' (override with `.groups` argument)
```

```
## # A tibble: 15 x 3
## # Groups:   X_incomg [5]
##    X_incomg                      renthom1           count
##    <fct>                         <fct>              <int>
##  1 Less than $15,000             Own                21485
##  2 Less than $15,000             Rent               25166
##  3 Less than $15,000             Other arrangement   4951
##  4 $15,000 to less than $25,000  Own                44378
##  5 $15,000 to less than $25,000  Rent               27306
##  6 $15,000 to less than $25,000  Other arrangement   3999
##  7 $25,000 to less than $35,000  Own                33322
##  8 $25,000 to less than $35,000  Rent               13224
##  9 $25,000 to less than $35,000  Other arrangement   1866
## 10 $35,000 to less than $50,000  Own                47081
## 11 $35,000 to less than $50,000  Rent               12228
## 12 $35,000 to less than $50,000  Other arrangement   1720
## 13 $50,000 or more               Own               157503
## 14 $50,000 or more               Rent               19164
## 15 $50,000 or more               Other arrangement   3220
```

```
ggplot(data=q2_data)+
  geom_bar(aes(x= renthom1, fill = X_incomg))+
  facet_wrap(~ X_educag ,ncol = 1)+
  xlab("Income Level")+
  guides(fill = guide_legend("Income Level"))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))
```



Insights: The above data shows that there is clearly a relationship between Education Level, Income, and Home Ownership. Although the relationship between Education Level, and Income may be expected, I was more surprised about the Education Level and Home Ownership relationship. From our last graph, we can see that Home Ownership Proportions are higher for those that have completed college than those who have not. This leads me to believe there may be a third variable, such as loan acceptance rates that may be confounding this relationship.

This would be a great topic of study to possible gather more data in and try to generalize this relationship.

Research question 3: Restating the Question from Above:

Do respondents that wear their seat belt and have gotten a flu shot visit the doctor more than respondents that do not fit that criteria?

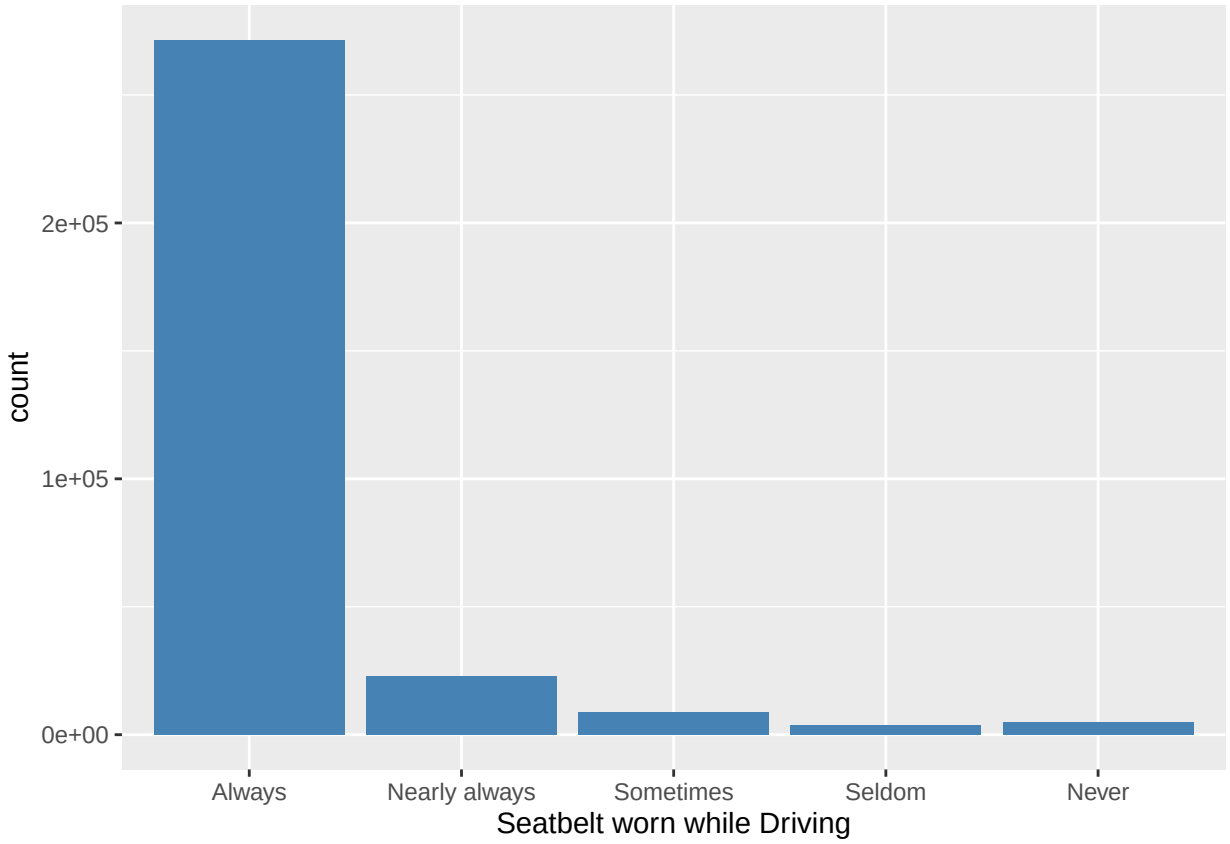
First lets discuss the variables that we will be using for this question. The seat belt use is a ordinal categorical variable, flu shot variable tells us if the respondent has gotten a flu shot in the last 12 months, and the doctor visits variable tells us how many visits a person has made to the doctor over the past 12 months.

Lets extract and clean our data. The good thing for us here is that in our entire data set, as we have seen above, all we have had to do is remove NA values. We haven't had the need to fill in missing values or anything else of the sort. However, this time, we want to get rid of some extra variable values that we would not like to consider.

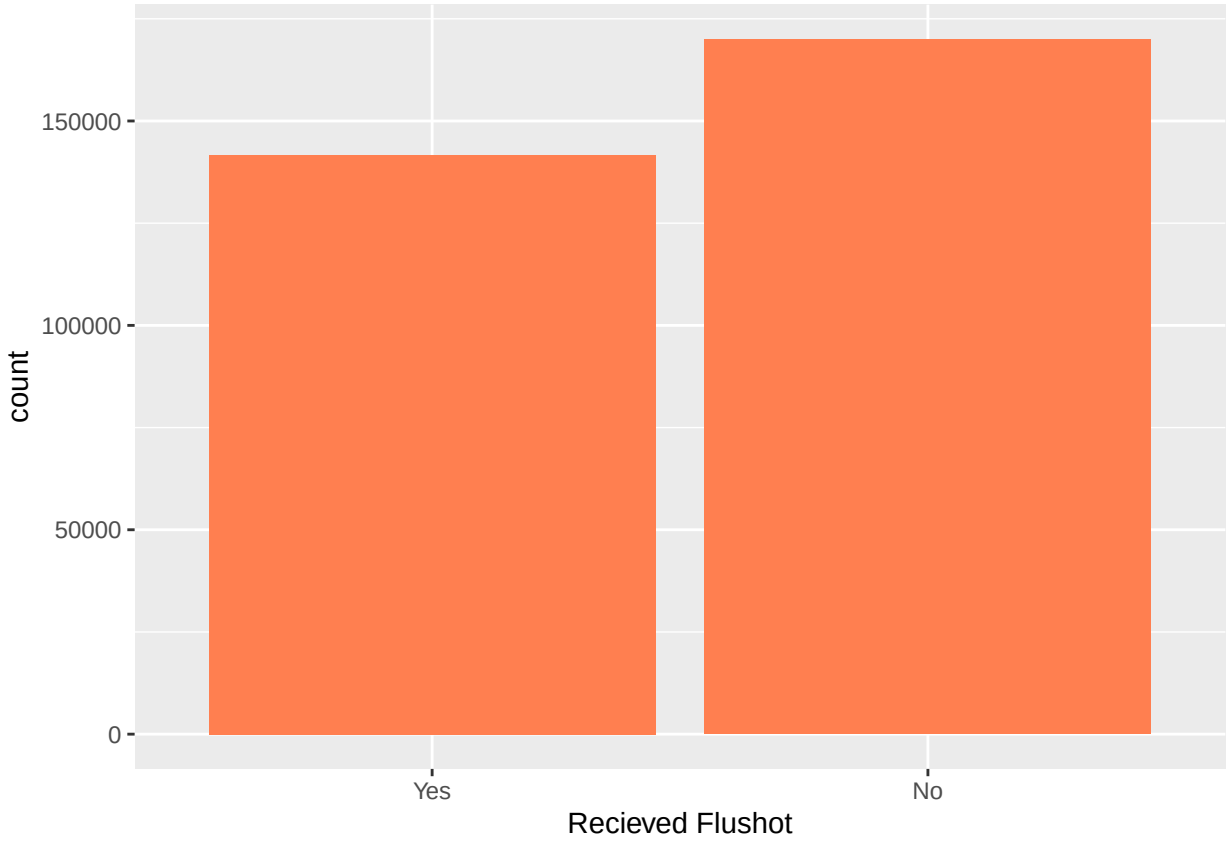
```
q3_data <- brfss2013 %>%
  select(seatbelt,drvisits,flushot6) %>%
  filter_at(vars(seatbelt,drvisits,flushot6),all_vars(!is.na(.)))
q3_data <- q3_data %>%
  filter(seatbelt != "Never drive or ride in a car")
```

I want to start by looking at the distributions of all of the variables in question.

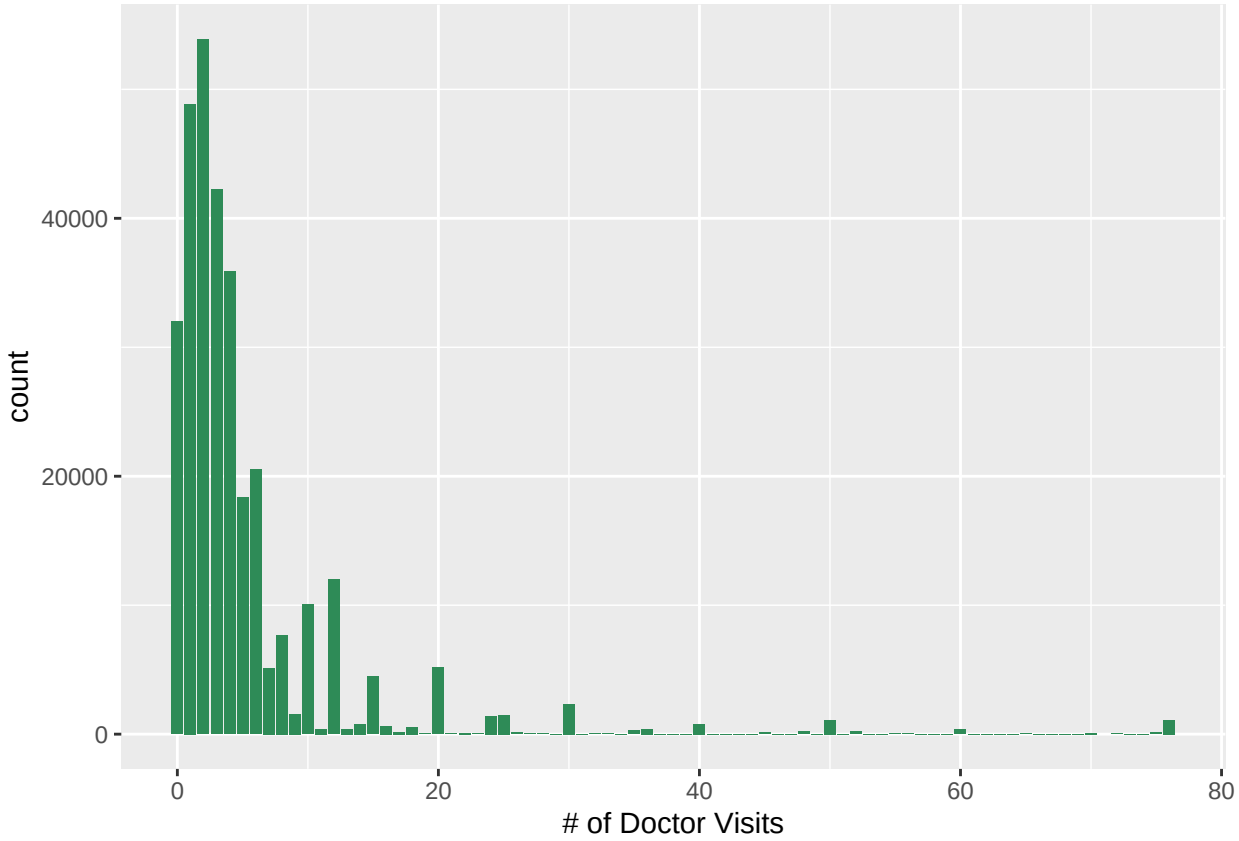
```
ggplot(data = q3_data)+
  geom_bar(aes(x= seatbelt), fill = "steelblue")+
  xlab("Seatbelt worn while Driving")
```




```
ggplot(data = q3_data)+
  geom_bar(aes(x= flushot6), fill = "coral")+
  xlab("Recieved Flushot")
```



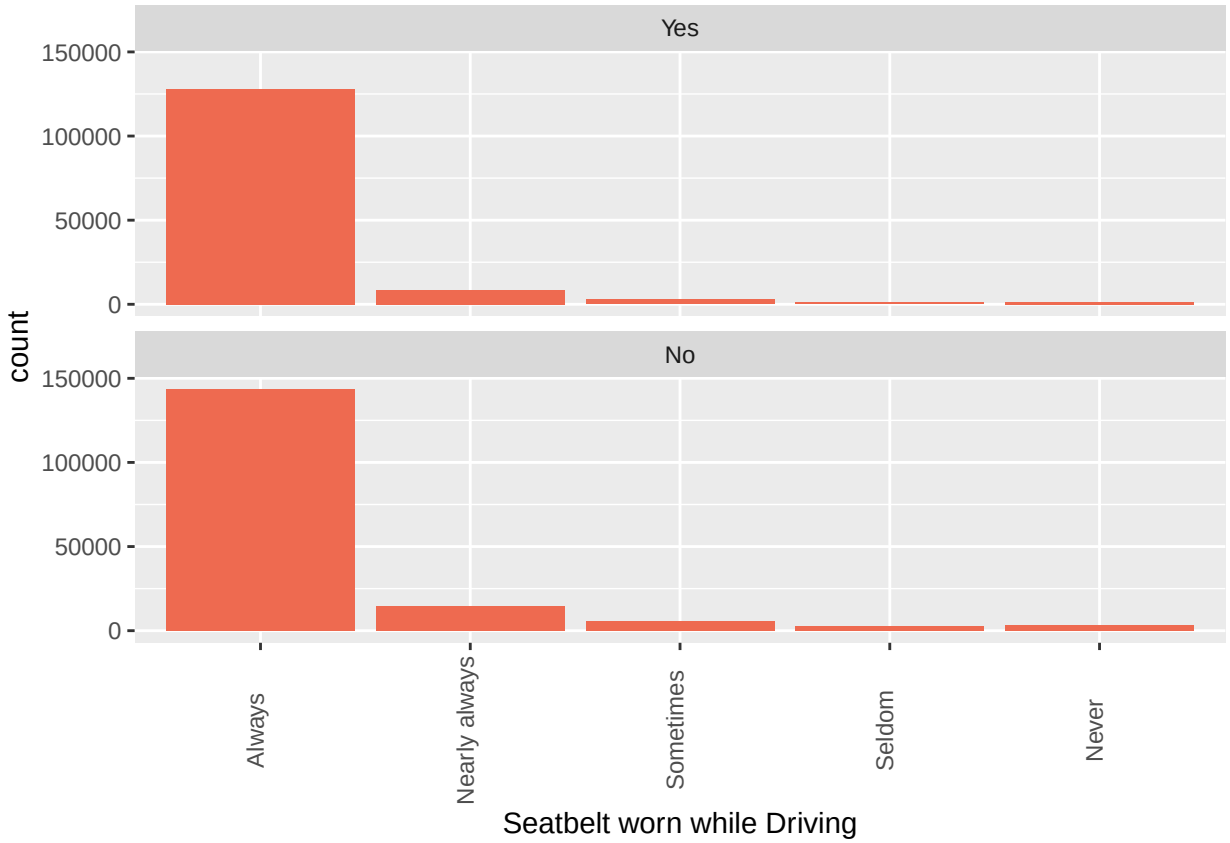

```
ggplot(data = q3_data)+
  geom_bar(aes(x= drvisits), fill = "seagreen")+
  xlab("# of Doctor Visits")
```



We can see from the above that, that overwhelmingly people say that they Always wear their seatbelts, More people have not gotten their flu shot than those who have, and the fistribution for doctor visits is right skewed with a peak about 3 visits.

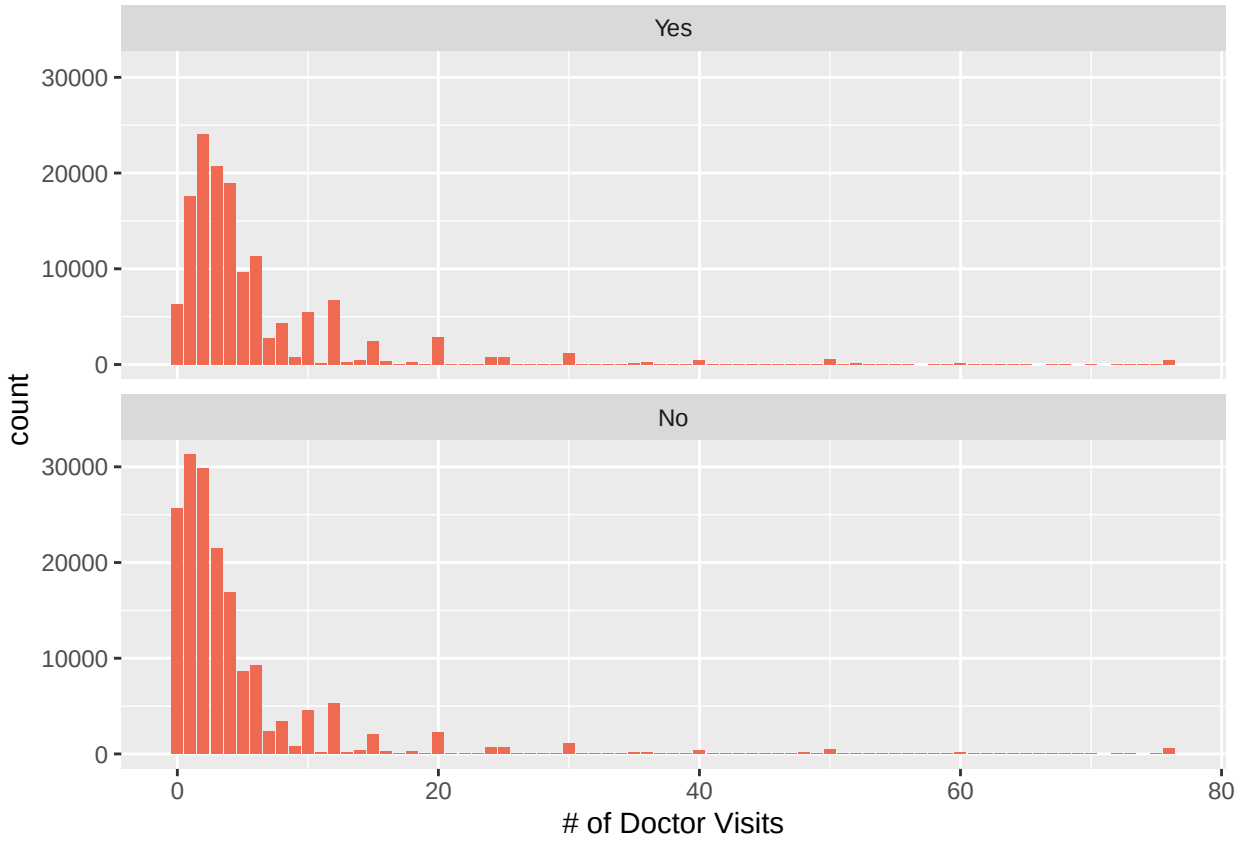
Lets dive more into the relationships between these variables.

```
ggplot(data = q3_data)+
  geom_bar(aes(x= seatbelt), fill = "coral2")+
  facet_wrap(~ flushot6, ncol = 1)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.2))+
  xlab("Seatbelt worn while Driving")
```



These distributions are similar, so it seems that having gotten your flushot does not relate to the frequency of wearing your seat belt.

```
ggplot(data = q3_data)+
  geom_bar(aes(x= drvisits), fill = "coral2")+
  facet_wrap(~ flushot6, ncol = 1)+
  xlab("# of Doctor Visits")
```



Again, there doesn't seem to be a difference between the data set distribution based on the flu shot variable alone. But there does seem to be a shift towards the right in terms of its peak value for the group that have gotten their flu shot. We can look at a table to have a clearer view.

```
q3_data%>%
  group_by(flushot6) %>%
  summarise( meanDR= mean(drvisits), medDR = median(drvisits))
```

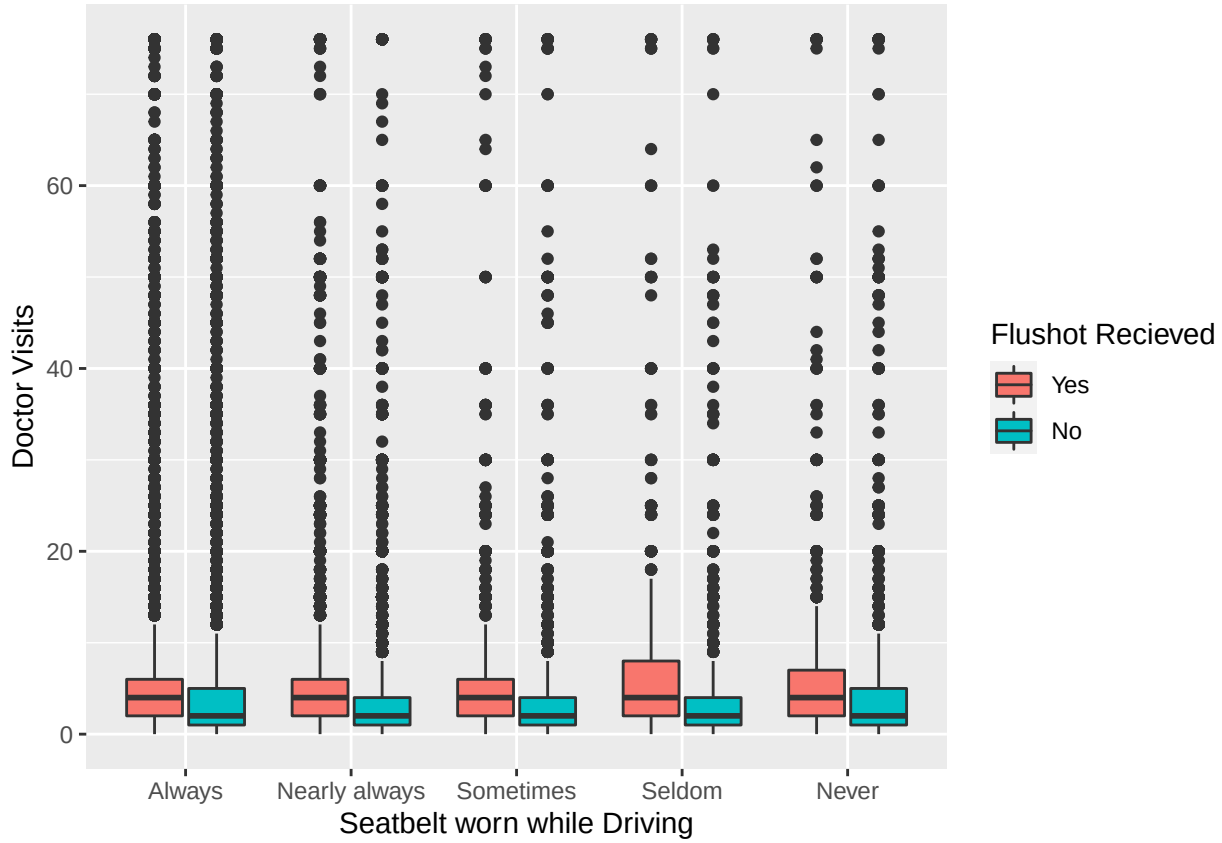
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   flushot6 meanDR medDR
##   <fct>     <dbl> <dbl>
## 1 Yes        6.10     4
## 2 No         4.59     2
```

Here we can see that the Mean and Median are both higher for the Doctor Visits variable, for the group that has gotten their flu shot.

We can now try and see if these three variables combined have a relationship that may be of interest to us. We will be generating a box plot

```
ggplot(data = q3_data, aes(x= seatbelt, y= drvisits, fill = flushot6)) +
  geom_boxplot()+
  xlab("Seatbelt worn while Driving")+
  ylab("Doctor Visits")+
  guides(fill = guide_legend("Flushot Recieved"))
```



Insights:

From all of the data visualizations and analysis above, I think one of the more interesting questions to study would be if health consciousness can be predicted by using variables like Flu Shot. If that is the case we might be able to draw a relationship between many more interesting things such as health conditions and likelihood of recieving preventative medicine.