Project 1 by Darien Lizano

Welcome to Revolving Door Consulting!

First and foremost, I would like to thank the Foundation for the Future of Work and Society for choosing our firm to conduct further analysis for your survey research. Although there were some challenges to visualize aspects of the dataset, our firm managed to overcome these through collaboration and hard work. In this report, our firm will guide you through all the steps taken to ensure that you have a full understanding of what we did, how we did it, and why we did it.

Loading Packages

To start, we must ensure that we have all neccessary packages installed and/or loaded in to begin our work. Listed below are all packages that were used for our analyses and visualizations.

- Dpylr is used for summarizing data to receive statistics
- readr is used to load in the provided survey dataset
- ggplot2 is used for mapping our visualizations
- viridis and RColorBrewer are used for color patterns in our visualizations
- ggthemes is used for the themes of our visualization
- ggrepel is used for inserting text into our visualizations

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Warning: package 'viridis' was built under R version 4.0.4

## Loading required package: viridisLite

## Warning: package 'ggthemes' was built under R version 4.0.4

## Warning: package 'ggrepel' was built under R version 4.0.4
```

Loading Dataset

Here, we are loading in the provided survey dataset. This dataset contains data from a survey that was conducted by the Foundation for the Future of Work and Society, involving a number of work and family variables from a relatively large sample. Once the dataset is loaded in, we save it as an object named "gssProject1."

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   X1 = col_double(),
##   year = col_double(),
##   id = col_double(),
##   age = col_double(),
##   householdSize = col_double(),
##   hrsUsuallyWork = col_double(),
##   hrsSpouseUsuallyWork = col_double(),
##   numChildren = col_double(),
##   hrsEmailPerWeek = col_double(),
##   hrsWWWperWeek = col_double()
## )
## i Use `spec()` for the full column specifications.
```

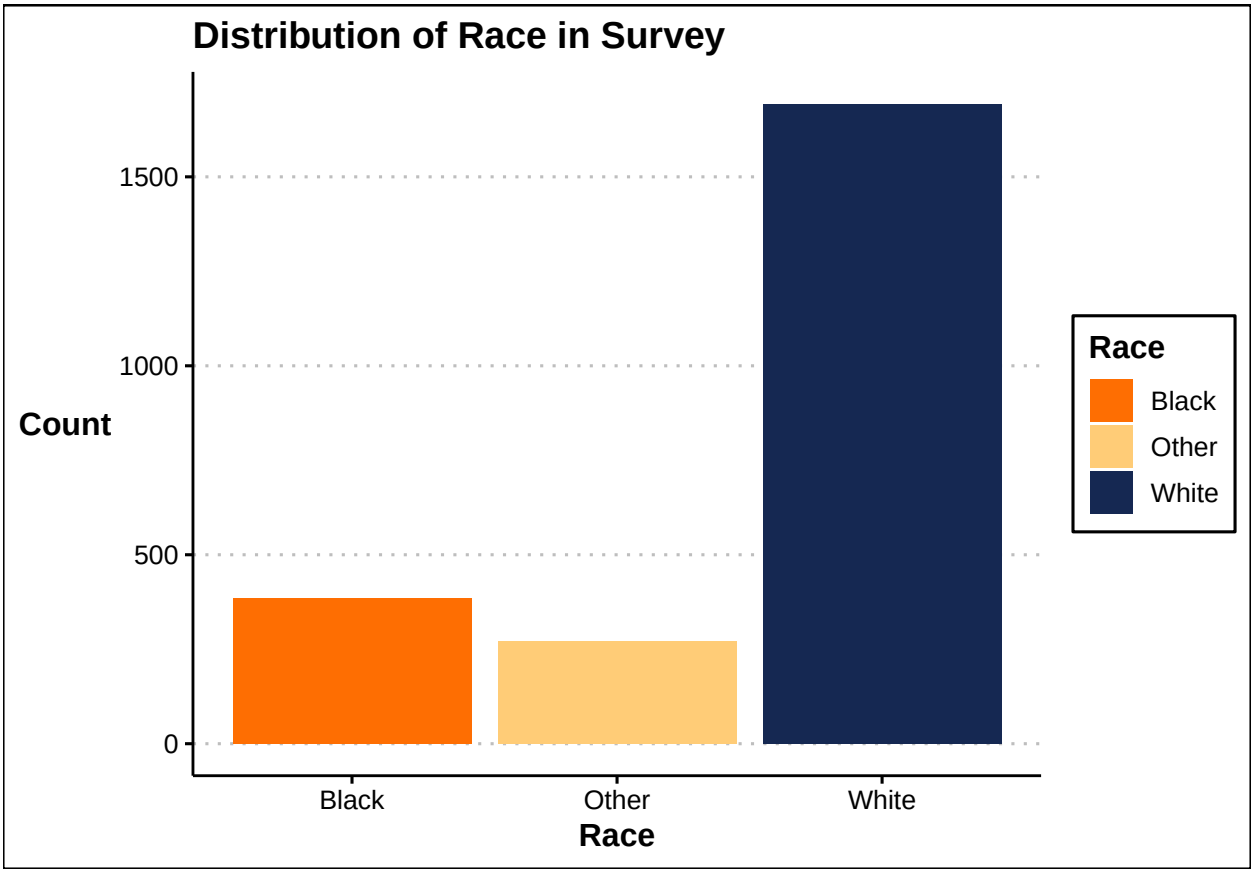
Question 1: What was the final sample size and what is the general demographic composition of the sample?

To figure out the sample size, we used the glimpse() function that gives us the total number of entries in the dataset, as well as the variable names and the types of variables that are contained in the dataset. From the provided dataset, the final sample size is 2,348 recipients. A caveat to this number is that there were many variables that contained missing data; for this, there were new datasets created, tailored to the specific questions asked by your organization.

For us to understand what the general demographic is for this survey, we needed to decide which variables we wanted to test. Our firm selected the variables 'race', 'sex', 'age', and 'marital status'. We are confident that these chosen variables represent the general demographic.

Race

For the first variable, 'race', we graphed it as a vertical bar graph. The x-axis contains the three choices of race provided in the survey, "White," "Black," or "Other," and along the y-axis of the graph is the count (in other words, the amounts for each input of race), and the legend contains the three categories of race.

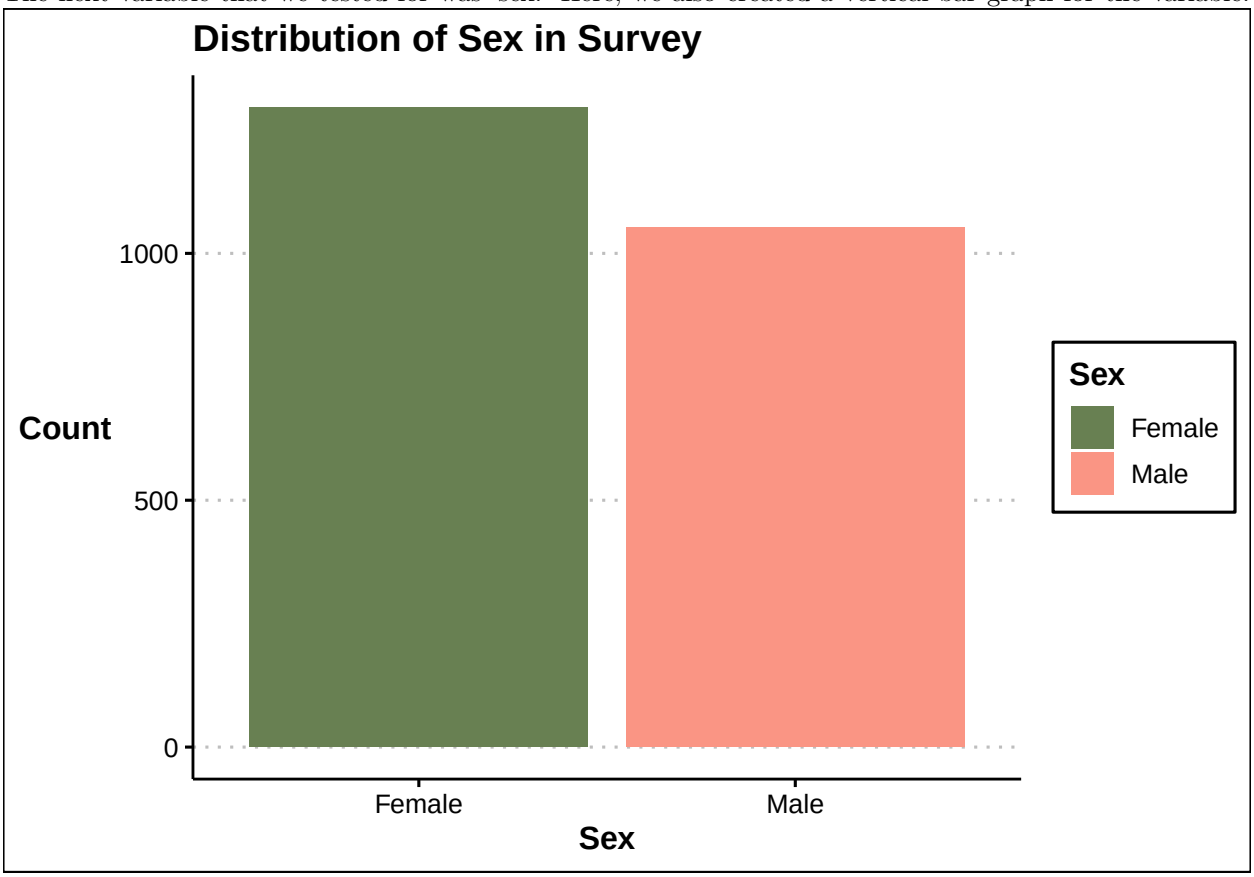



```
##
## Black Other White
##   385   270  1693
```

The majority of the survey takers identified themselves as White (1,693 recipients). The second largest group were those who identified themselves as Black (385 recipients), and the third group idenfitied themselves as Other (270 recipients).

Sex

The next variable that we tested for was 'sex.' Here, we also created a vertical bar graph for the variable.



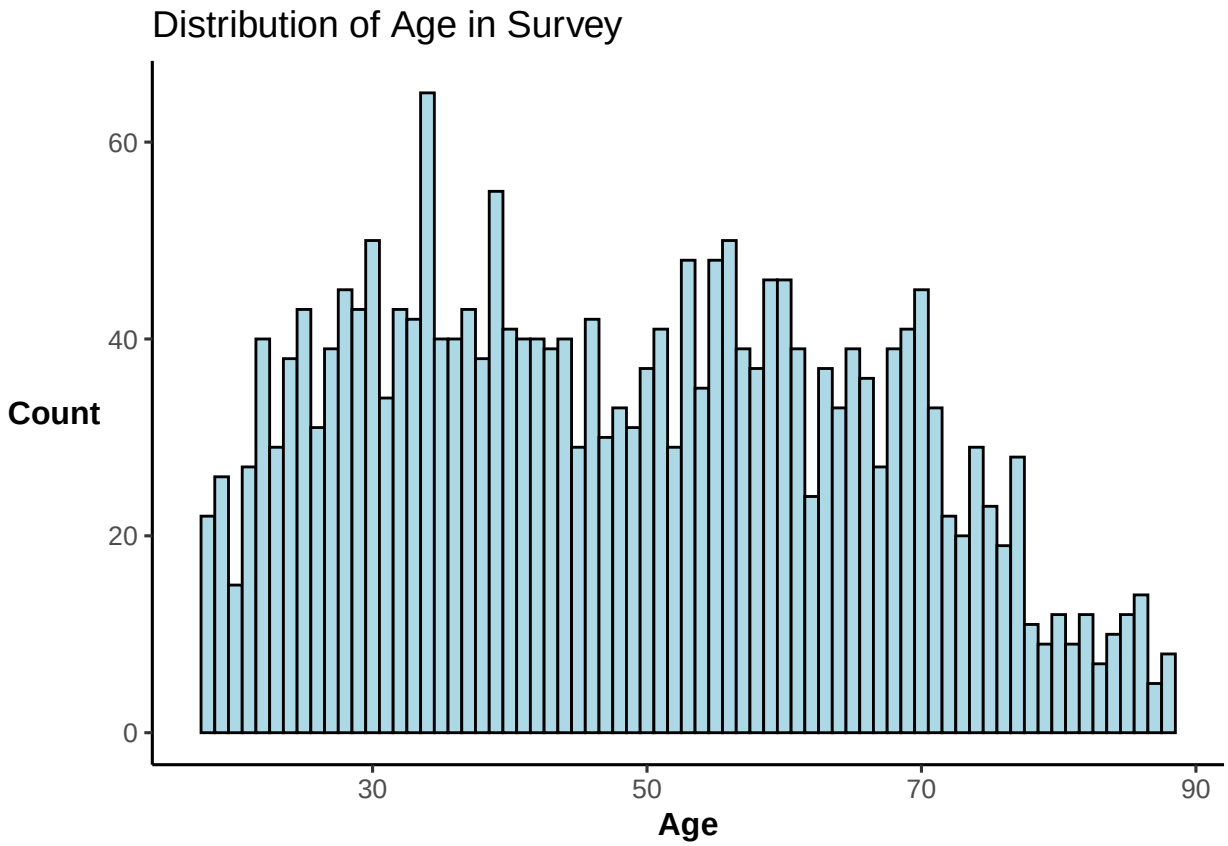

```
##
## Female   Male
##   1296   1052
```

This graph shows us there were more Females (1,296) who took this survey compared to Males (1,052). We kept count on the y-axis, changed the x-axis for it to contain the sex variable, and our legend contains sex that are color coded.

Age

The next variable is 'age', which is on the x-axis, and the y-axis remains as the count. For this variable, we determined that it would be best to used a Histogram to show your organization the distribution of age within the dataset.

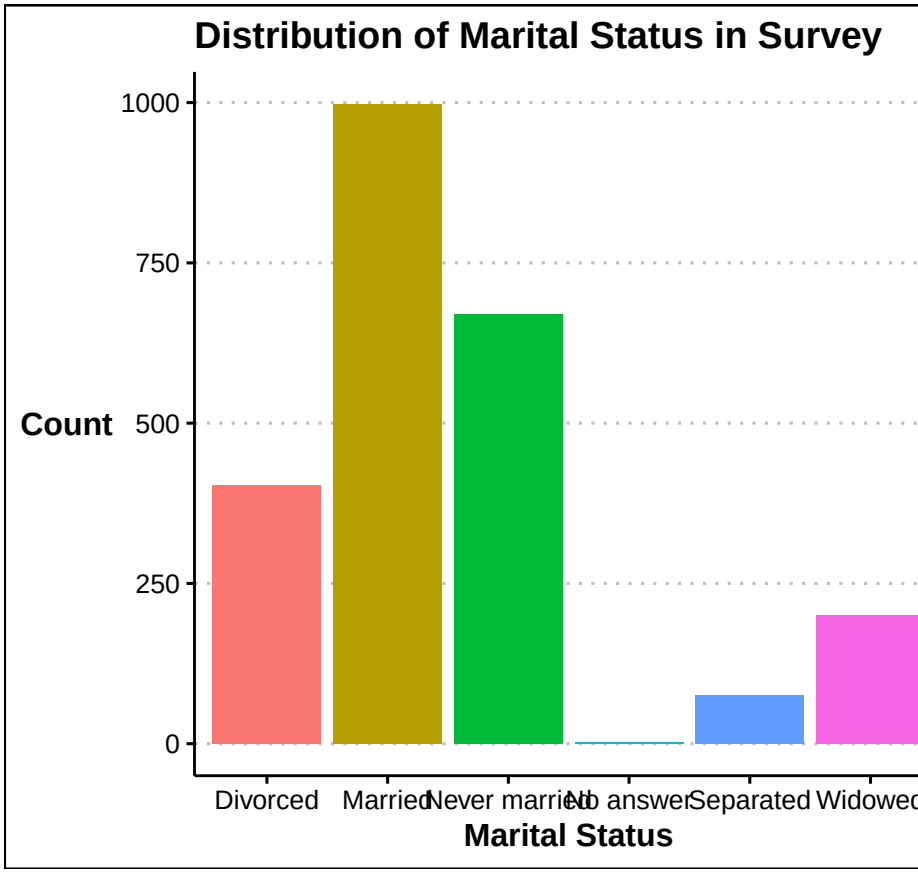
```
## Warning: Removed 36 rows containing non-finite values (stat_bin).
```



As we can see in this graph, the age groups display a normal distribution. The smallest age group is individuals counted that are 73 years and older. The largest age group within the dataset appears to be individuals between the ages of 30 and 40.

Marital Status

Lastly, the last variable we visualized was 'marital status'. Here we created a vertical bar graph that shows the distribution of marital stats groups within the survey, having marital status on the x-axis, count on the y-axis,



and our color-coded legend for martial status.

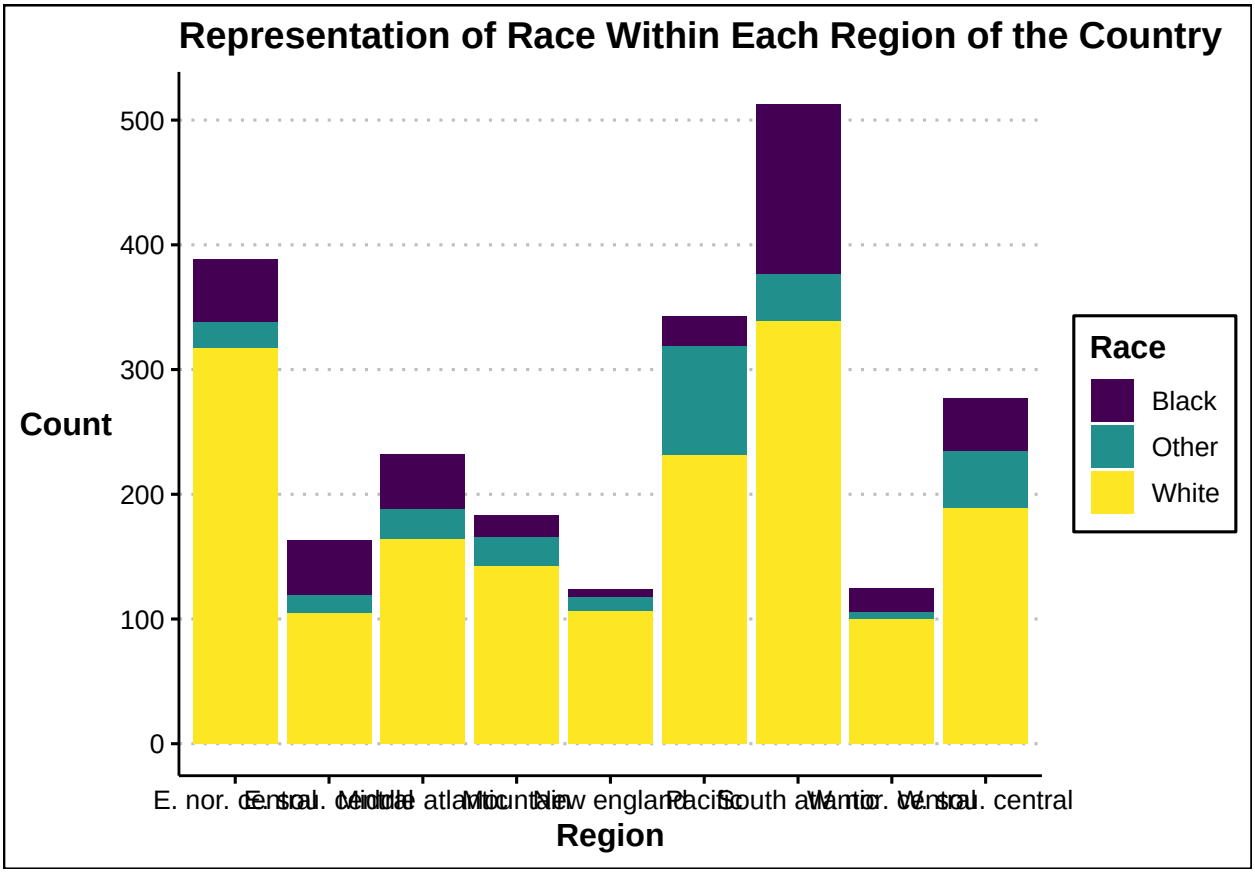
```
##
##      Divorced       Married Never married     No answer     Separated
##           403           998           670             2            75
##       Widowed
##           200
```

Those who are married were the biggest group in the survey (998 recipients), followed by those who were never married (670 recipients), those who are divorced (403 recipients), those who are widowed (200 recipients), those who are separated (75 recipients), and those who declined to answer (2 recipients)

From our visualizations, it is clear that the general demographic of this survey is a mix of many groups. The predominate groups of this survey are white females, within the ages of 30-40, who are currently married.

Question 2: Does the sample include a fair representation of race within each region of the country?

Our firm thought that for this visualization, we would stick with a vertical bar graph, but we decided to stack it to show the distribution of race within each region (located on x-axis), as the variable 'race' is color coded in



the legend.

```
##                  race
## region            Black Other White
##   E. nor. central    50    21   317
##   E. sou. central    44    14   105
##   Middle atlantic    44    24   164
##   Mountain           17    24   142
##   New england         6    12   106
##   Pacific            24    88   231
##   South atlantic    137    37   339
##   W. nor. central    20     5   100
##   W. sou. central    43    45   189
```

As we observe this graph, we are told that the majority of the recipients in this survey are located in the South Atlantic region of the country (a total of 513 recipients). There is no surprise that the race that is mostly represented is White across all regions. Our firm is reluctant, however, to claim that this sample contains a fair representation of race across all regions, due to the skew in data that is in favor of those recipients that identify as White.

Question 3: Is the number of hours usually worked in a week related to any of the following: trouble sleeping, job stress, job satisfaction, health, and happiness?

For this question, it was important for our firm to create new objects that contained the relevant variables we were testing for.

Hours Usually Worked and Trouble Sleeping

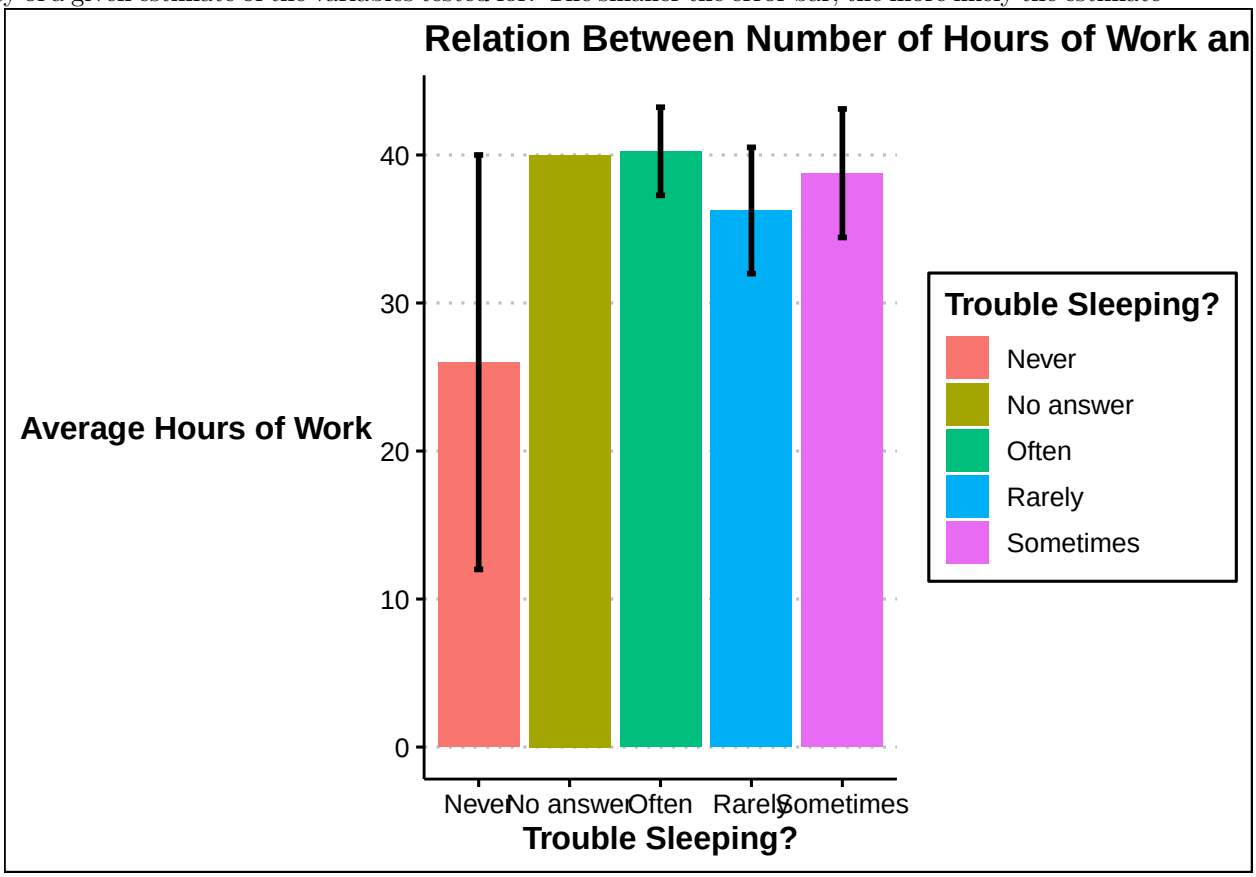
Our first object named question3_sleeping, contains the variables 'hrsUsuallyWork' and 'troubleSleeping.' For our objects, we made sure to filter through the selected variables to ensure we were getting the appropitate data to be tested for, without any errors or missing values. We used the summarize function to create new variables within our object so that we were able to graph the error bars. This allows us to view the average hours spent at their jobs, along with other statisical metrics. Using the filter(!is.na()) function allows us to remove missing values from our new object. It is important to note that we performed similar steps of filtering our objects in future code chunks to ensure a similar quality of data.

```
## Warning in qt(0.975, df = n - 1): NaNs produced
```

After creating the question3_sleeping object, we called for the object to verify that the new object contains relevant data, and that it is free from any missing values (in which it was free from missing values).

It is important to note that the response "No Answer" yielded missing values. For this, we ignore this variable for its statistical properties, but it remains in the object for visualization purposes.

After we verified our dataset, we continued onto our visualization of our new object. For this, we chose a vertical bar graph with error bars. On the x-axis are the categories for the trouble sleeping variable with the responses color coded within the legend, and on the y-axis is the average hours worked per week. These error bars represent the standard error that was calculated within our filtering of the dataset. This provides the uncertainty of a given estimate of the variables tested for. The smaller the error bar, the more likely the estimate is true.



Here, we see that the "Never" response contains the biggest standard error (14.0). The "Often" response yields a standard error of 2.976, which is the lowest of all other responses. This group also works the highest amount of average hours per week (40.25 hours). This gives us the confidence to say that those who say they often suffer from difficulty of sleeping due to the amount of hours worked per week, on average. This

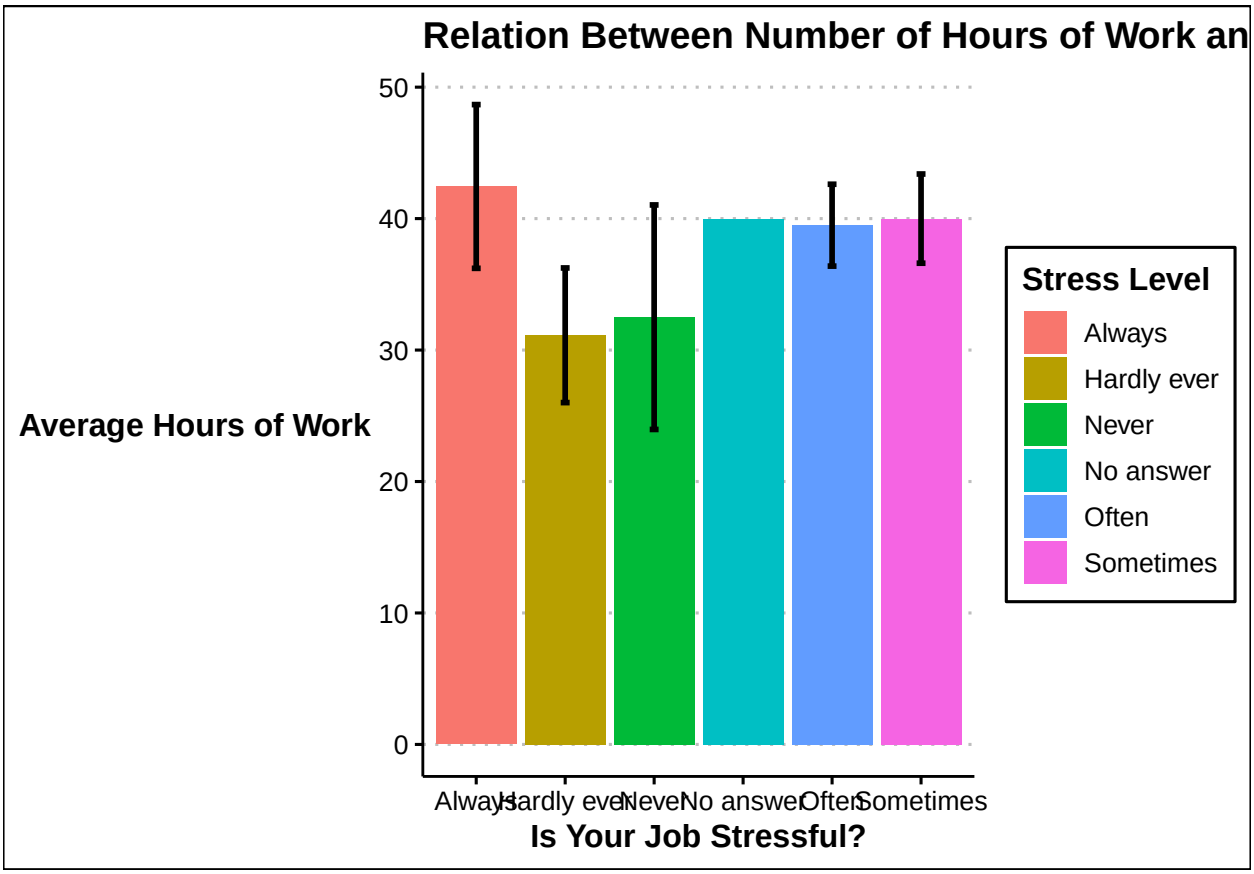
also means that the number of hours usually worked per week is related to how you sleep at night (if you are able to sleep without difficulty, that is).

Hours Usually Worked and Job Stress

The next variable used with 'hrUsuallyWork' is 'jobStress'. We created a new object named question3_stress, following the same steps as the previous object.

```
## Warning in qt(0.975, df = n - 1): NaNs produced
```

It is important to note that the response "No Answer" yeilded missing values. For this, we ignore this variable for its statisical properties, but it remains in the object for visualiation purposes.



Here, we used the same graphing method as we did for trouble sleeping. The only changes are the x-axis and legend which both contain the jobStressful variable. In this graph, we can see that the standard errors yielded a much lower variance than the trouble sleeping variable. We also see that the group that answered "Always" worked the most amount of hours, on average (42.44 hours). The group with the most responses answered as "Sometimes" (16 recipients) and they worked the second highest amount of hours, on average (40 hours). We are confident to say that a relation exists between the numbers of hours worked on average, and the amount of stress that someone experiences from their job.

Hours Usually Worked and Job Satisfaction

Next, we compare hours usually worked to 'jobtatisifaction'. Again, we peform the same filtering methods for our new object, question3_jobsatis, as our previous objects.

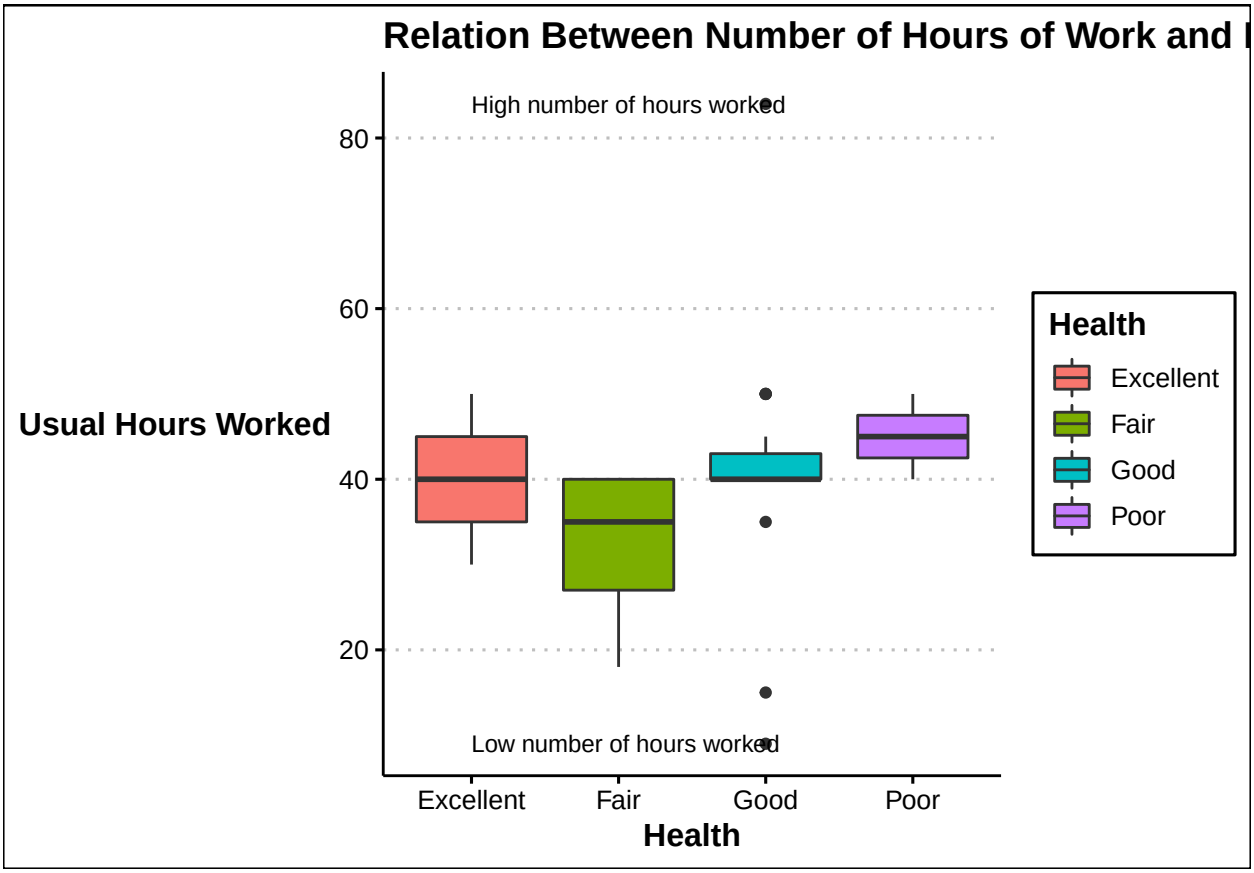
After verifying our object, we move on to graphing. We chose a vertical bar graph with error bars. On the x-axis is job satisfaction, on the y-axis is the count, and the legend is job satisfaction.

After we analazyed our object and graph, we were able to draw some conclusions. Excluding "No Answer" response, the "Not At All Satisfied" response was lowest (3 recipients). The highest response was "Somewhat Satisfied" (21 recipients). What was suprising is that those who are somewhat satisfied worked more hours than those who responded as not being at all satisfied in their job. This began to raise questions such as: what jobs do they have? Is it a job within their field of study, is it an entry level job that may lead them into their potential career, or does working more hours mean higher pay, therefore higher job satisfaction? Unfortunately, we were not given the data to answer these sorts of questions. What we are able to say is that there is a relationship between hours usually worked and job satisfaction, but it is difficult to pinpoint what it is.

Hours Usually Worked and Health

Our next variable paired with 'hrsUsuallyWork' is 'health'. For this object, we decided not to summarize with the piping operator since we are choosing a boxplot for this object. We still followed the same filtering methods, minus the summarize() function within the piping operator.

It is important to note that we did use the summarize() function for the new object, question3_health, for it is necessary to explain what the boxplot represents.

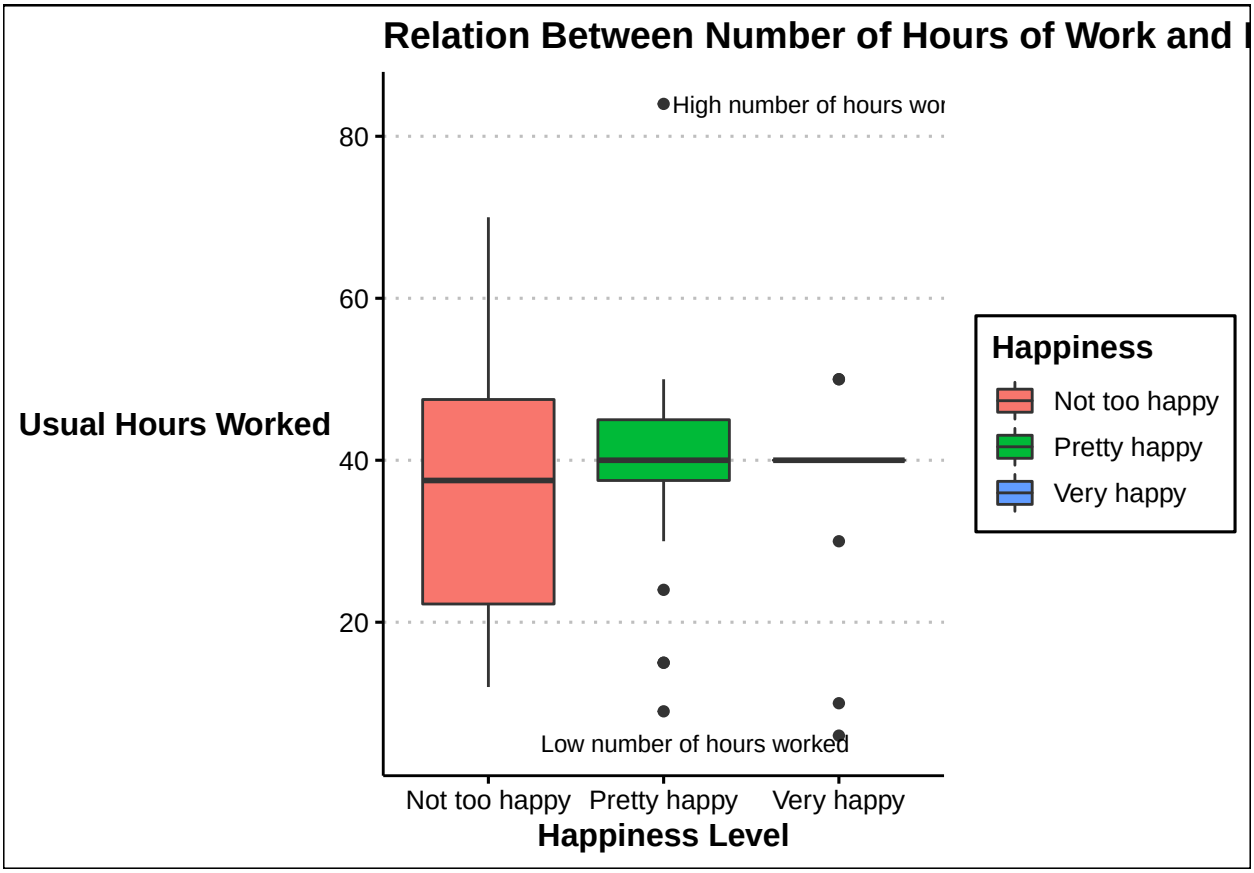


Looking at our summary of the object allows us to better understand the boxplot. Firstly, we will provide an overview of how to interpret a boxplot.The bottom horizontal line in the colored boxes represent the first quartile (the number between the smallest number and the median of the dataset), the middle horizontal line represents the median, the upper horizontal line that encloses the box is the third quartile (represents the number between the median and highest value), the vertical lines on the top and bottom of the boxes are the upper and lower whiskers (which represent the numbers outside of the middle 50% of the dataset), and the dots farther away from the boxes are outliers that represent data that falls outside of the normal distribution.

In our graph, we have health on the x-axis and in the legend, and on the y-axis is the usual hours worked variable. Looking at the boxplot, we notice that there are outliers that are labeled with text. These outliers represent hours worked that are far from the distribution of the data, having 84 hours as the maximum outlier, and 9 hours as the minimum. We can also see that those who responded, claiming that they have poor health, have the highest median of all other groups, but also contain the second lowest amount of recipients. Those who claim their health is in excellent condition have a median of 40 hours worked per week, and also have higher quartiles than all other responses. Those who responded that their health is in fair condition have the lowest median of all, and also the greatest lower quartile, meaning they have more respondents who worked less amount of hours. It is safe to say that there is a relation between the amount of hours worked and one's health condition. The effects on the person's health varies, as we see that those who reported poor health worked slightly more hours than those who said they had excellent health, and those who reported as having good health worked more hours than those who reported of having fair health.

Hours Usually Worked and Happiness

The next and final variable we compare 'hrsUsuallyWork' with 'happiness'. Since we chose to stick with a boxplot for this comparison as well, we conducted the same filtering process as the previous object.



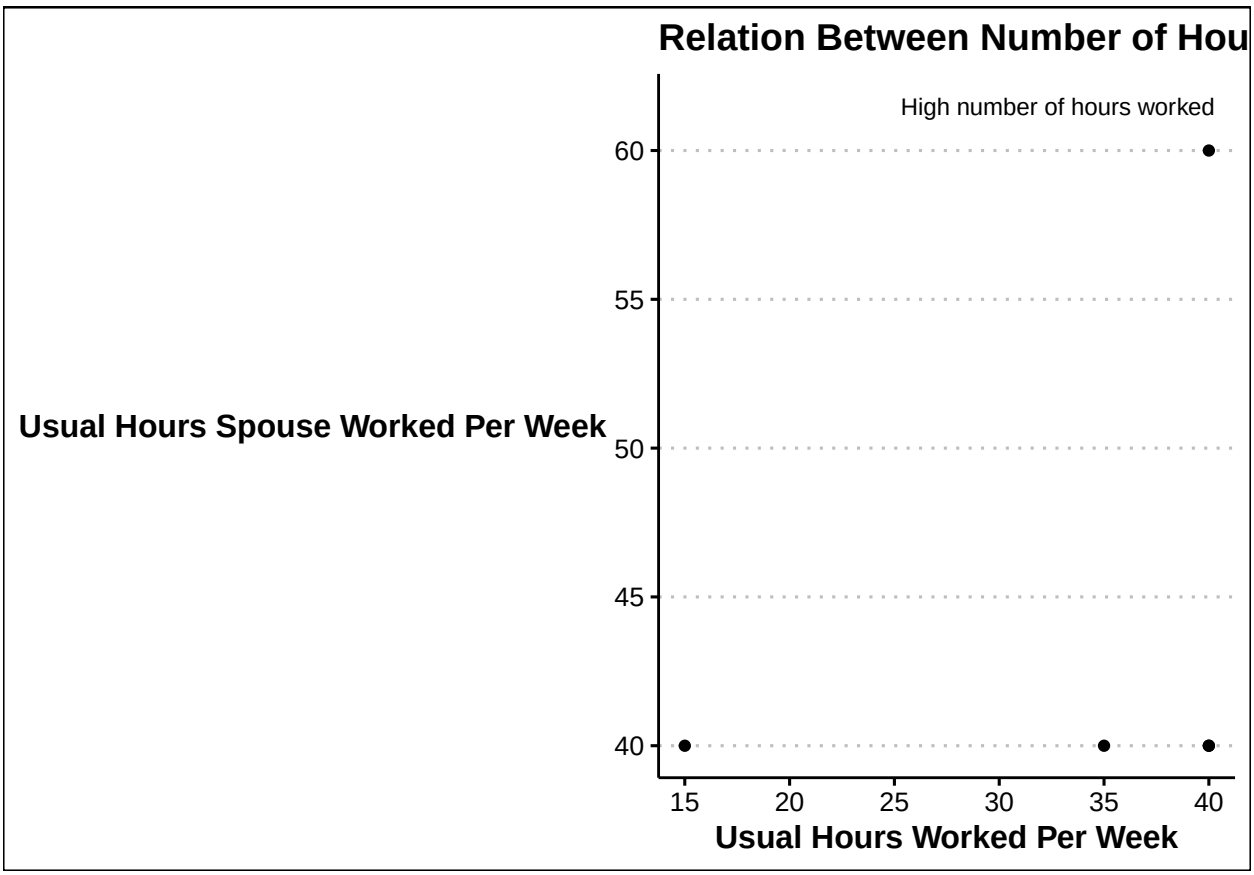
We placed happiness level on x-axis and in the legend.

As we look at this boxplot, we notice something very different. We see that those who responded "Very Happy" did not yeild a box. This occurred because those who responded to "Very Happy" had an exact average and median of 40 hours per week, along with a few outliers. The maximum value of the boxplot is 84 hours, which is the max outlier for "Pretty Happy", and the minimum value was 6 hours for the "Very Happy" resposne. The "Not Too Happy" response yielded the biggest box of all other responses, meaning that it contained various amounts of hours worked. We noticed here that the median is the lowest of all other responses. This gives us confidence to say there there is a relation that exists between hours worked and overall happiness. We can conclude that working slightly more hours can lead to lower happiness.

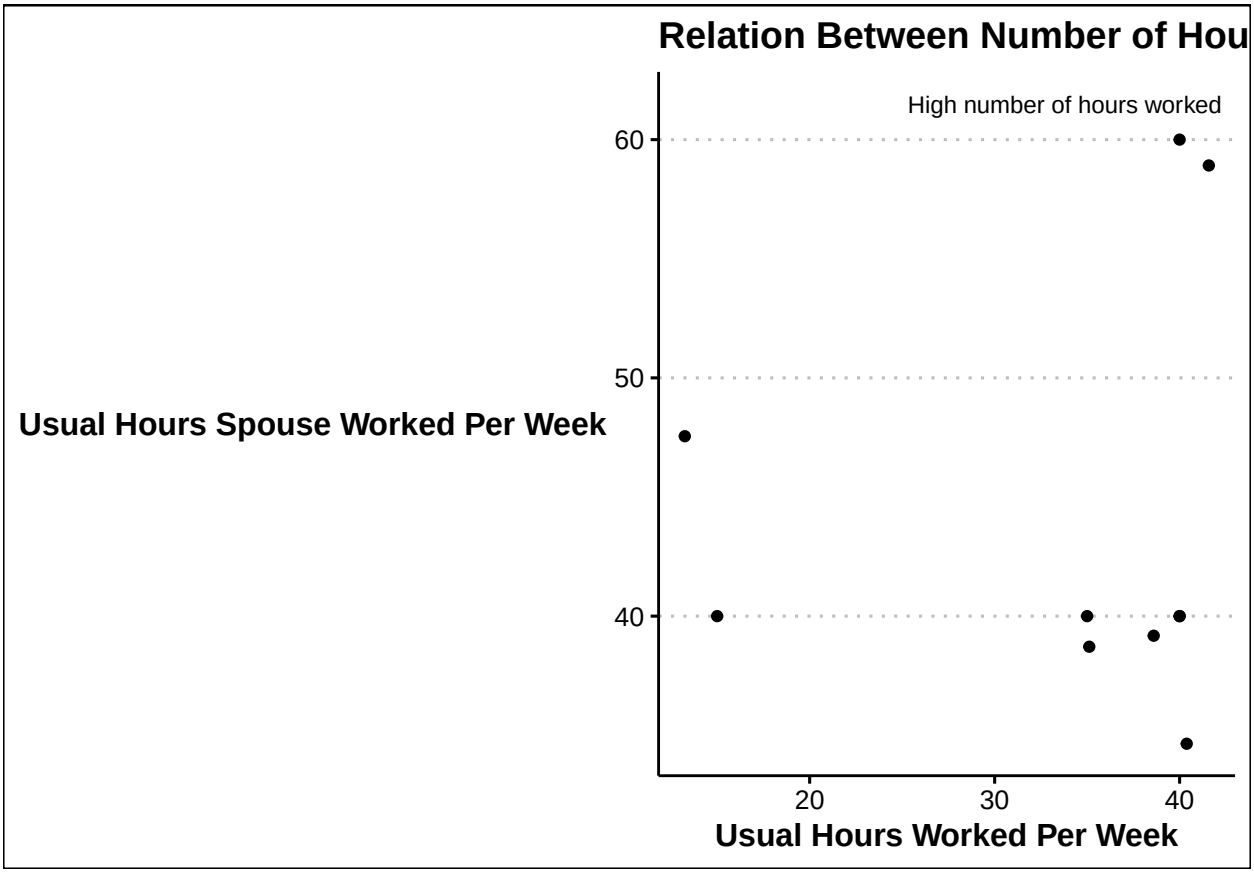
Question 4: What is the relationship between the number of hours spouses usually work each week? That is, as one spouse works more, does that mean the other spouse works less? Or, vice versa?

For this question, we followed the exact apporach as the previous object. Our new object, question4, contains the variables 'hrsUsuallyWork' and 'hrsSpouseUsuallyWork'.

For this object, we can see that hours that the spouse usually works are higher than hours usually worked by the surveyee. It is important to note that, by these value points, we will have overplotted points on our scatterplot. The following graph is a scatterplot wihtout using the geom_jitter() function. For both graphs, the usual hours that the spouse worked per week is on the y-axis, whereas the usual hours worked for the surveyee is on the x-axis.



The following graph is the same as the previous scatterplot, only this time we implement the geom_jitter() function.



The geom_jitter() function adds random variation to the points so that it creates a scattering effect, allowing for overplotted points to be shown. Using geom_jitter(), we see more points than the first scatterplot that was created for this question; this provides more insight as to what the graph is trying to tell us.

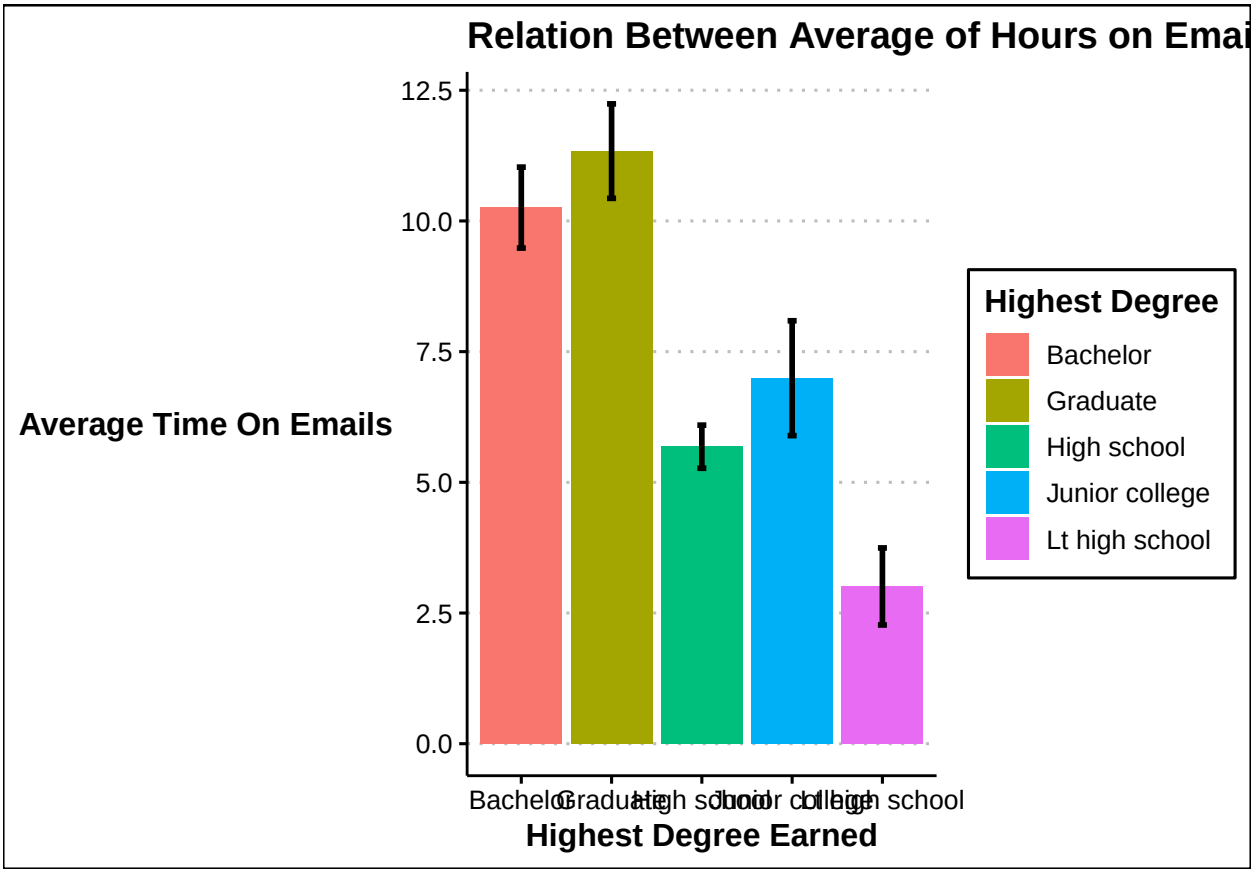
After looking at the graph, we can see that there is a relation between the amount of hours the surveyee worked compared to their spouse. The more hours the surveyee works, the closer those hours are to the hours worked by their spouse. It is important to note that there are two outliers, one for 15 hours worked and the other for 60 hours worked. These outliers are shown to be insignificant in this analysis.

Question 5: Does the amount of time that people spend on email each week vary as a function of the highest degree that they've earned? That is, do people with more education, have jobs where they spend more time on email?

For this question, we reverted back to the filtering method that uses the summarize() function, in order for us to plot error bars with our vertical bar graph. Our object for this is named question5.

By looking at this object, we see that the majority of the recipients have a High School degree as their highest degree obtained (719), followed by a Bachelors Degree (298), Graduate Degree (158), Less Than High School (123), and Junior College (121). We alo observed that those with a Bachelor's Degree average the most hours on their email. Let's see how this translates in our visualization.

For our graph, we chose a vertical bar graph; on the y-axis is the average hours spent on email, and on the x-axis and in the legend is the highest degree earned.



As we observe this visualization, we can confirm that those with a Graduate Degree spend the most time on their emails, while also having the second highest standard error (0.90), which further increases the confidence. Following are Bachelor's degree (10.25 hours), Junior College (6.99 hours), High School (5.68 hours), and less than High School (3.00). Therefore, we are confident in saying that there is a strong relation between their highest degree earned and the amount of time they spend on their emails per week; the higher the earned degree, the more time they spend on emails.

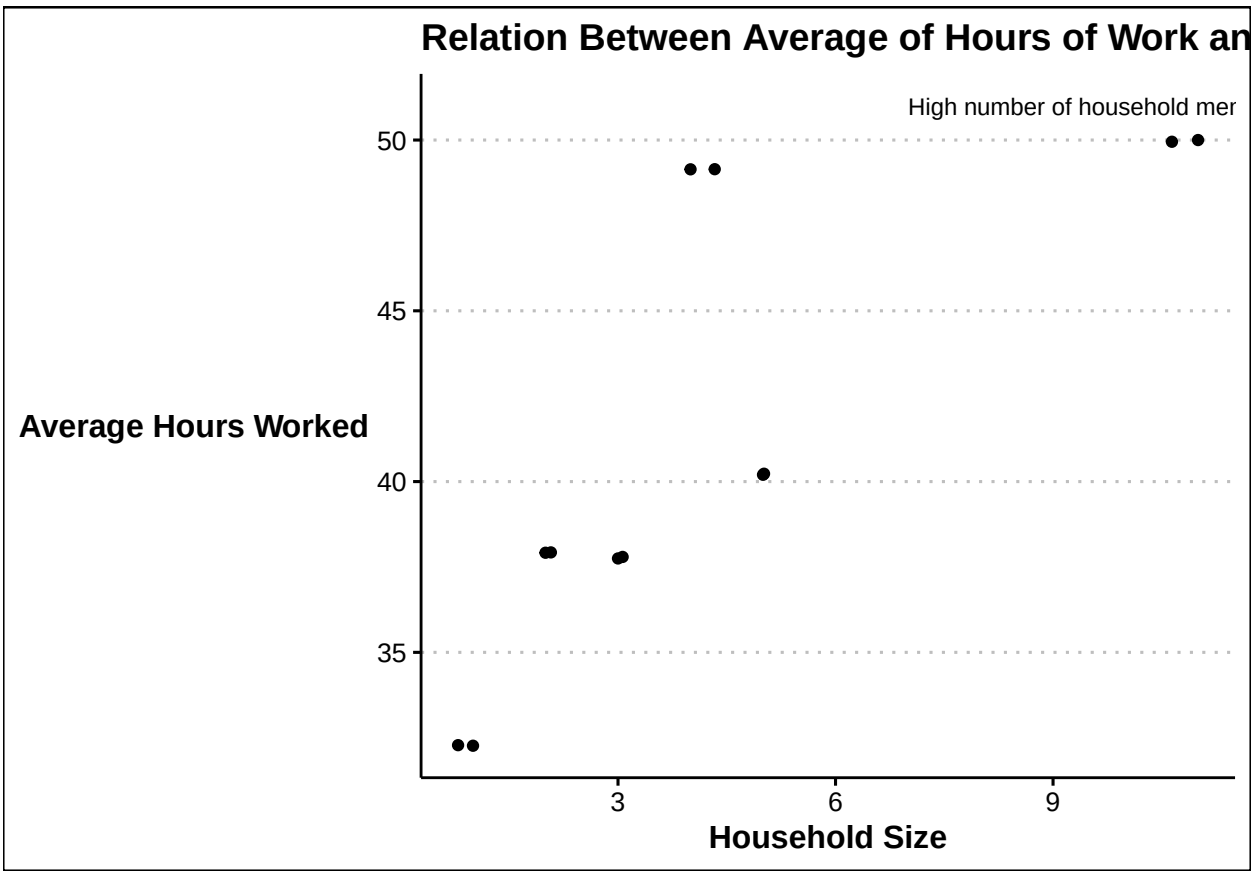
Question6: Do people who live alone work more than people who live with others in their household?

For this question, we filtered for our new object, question6, using the summarize function. Only this time, we were not plotting error bars, but the average hours worked in the week.

```
## Warning in qt(0.975, df = n - 1): NaNs produced
```

It is important to note that the response "11" for household size yeilded missing values. For this, we ignore this variable for its statisical properties, but it remains in the object for visualiation purposes.

After analyzing this object, we concluded that as the number of household members increases, so does the average hours worked per week; the only exception is for the household size = 5 response, where we notice a decrease in average hours worked compared to the household sizes of 4 and 11.



In this graph, household size is on the x-axis, and average hours worked is on the y-axis. For this scatterplot, we used the geom_jitter() function. Here, we observed a clear relation between household size and average hours worked. It is apparent that the bigger the household, the more hours that person works throughout the week. To answer this question: no, people who live alone do not work more than people who live with others.

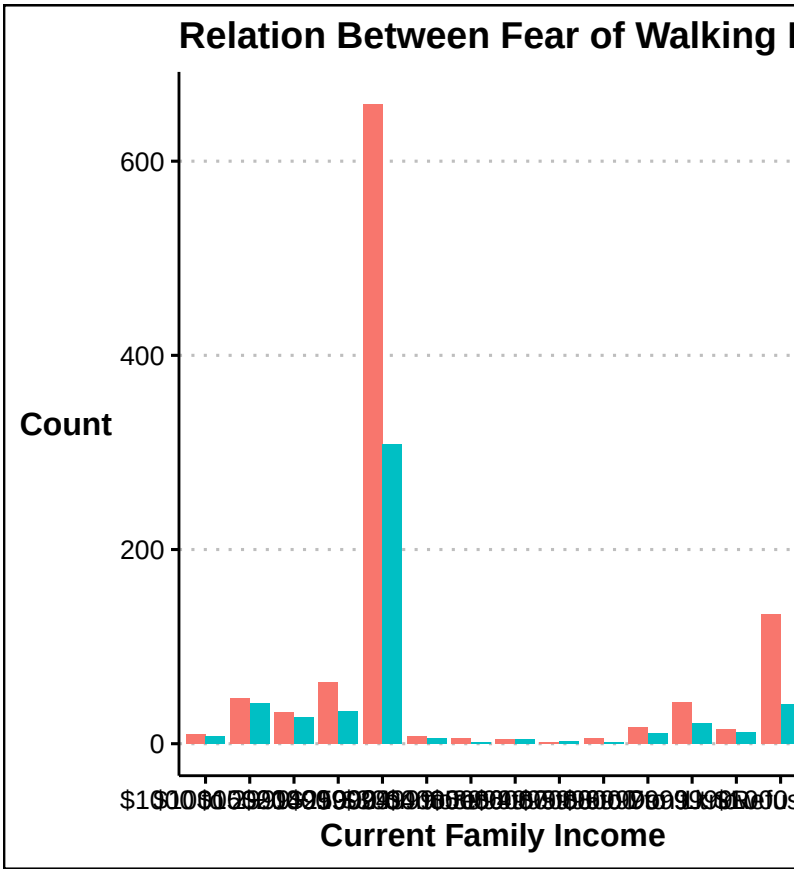
Question 7: Is fear of walking in one's own neighborhood at night related to any of the following: current family income, age, sex, race, or region?

The first step that we took in this question was to create the question7 object that filters out all unnecessary categorical levels other than the 'Yes' and 'No' responses within the 'afraidWalkNightNeighborhood' variable.

Afraid of Walking at Night Alone and Current Family Income

Once we filtered the object question7, we then used this object as the dataset used to create the objects needed for this question. The first object we made selected both the afraidWalkNightNeighborhood and 'currentFamIncome' variable, to determine if there is a relation between the two variables.

We determined that a vertical bar graph would be best to visualize this relation. We selected 'currentFamIncome' as our variable on the x-axis and legend, count on the y-axis, and used the fill= function for the re-

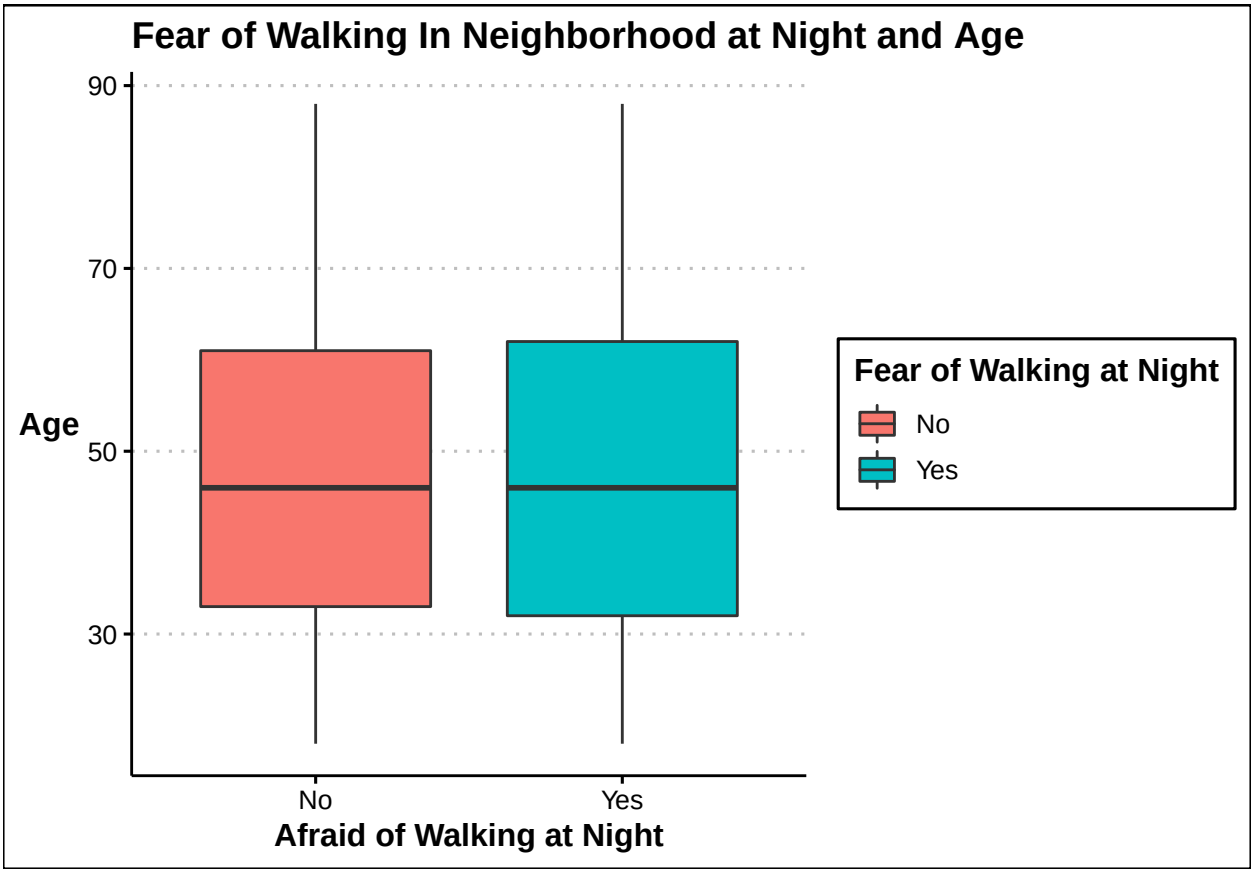


sponses to the 'afraidWalkNightNeighborhood' variable. As we observe the graph, it is clear that the majority of recipients live in households that currently make $25k or more per year as a family income; the majority said that they are not afraid to walk at night within their neighborhood. In fact, there is no income level where those who are afriad exceed those who arent. Therefore, we are confident to say that there is no relation between current family income and being afraid of walking at night within their neighborhood.

Afraid of Walking at Night Alone and Age

For this part of the question, similar steps were taken as with the previous graph. the only difference is that we decided to choose a boxplot for this variable. Due to the various amounts of different ages in the dataset. We have age on the y-axis, and fear of walking at night on the x-axis and legend.

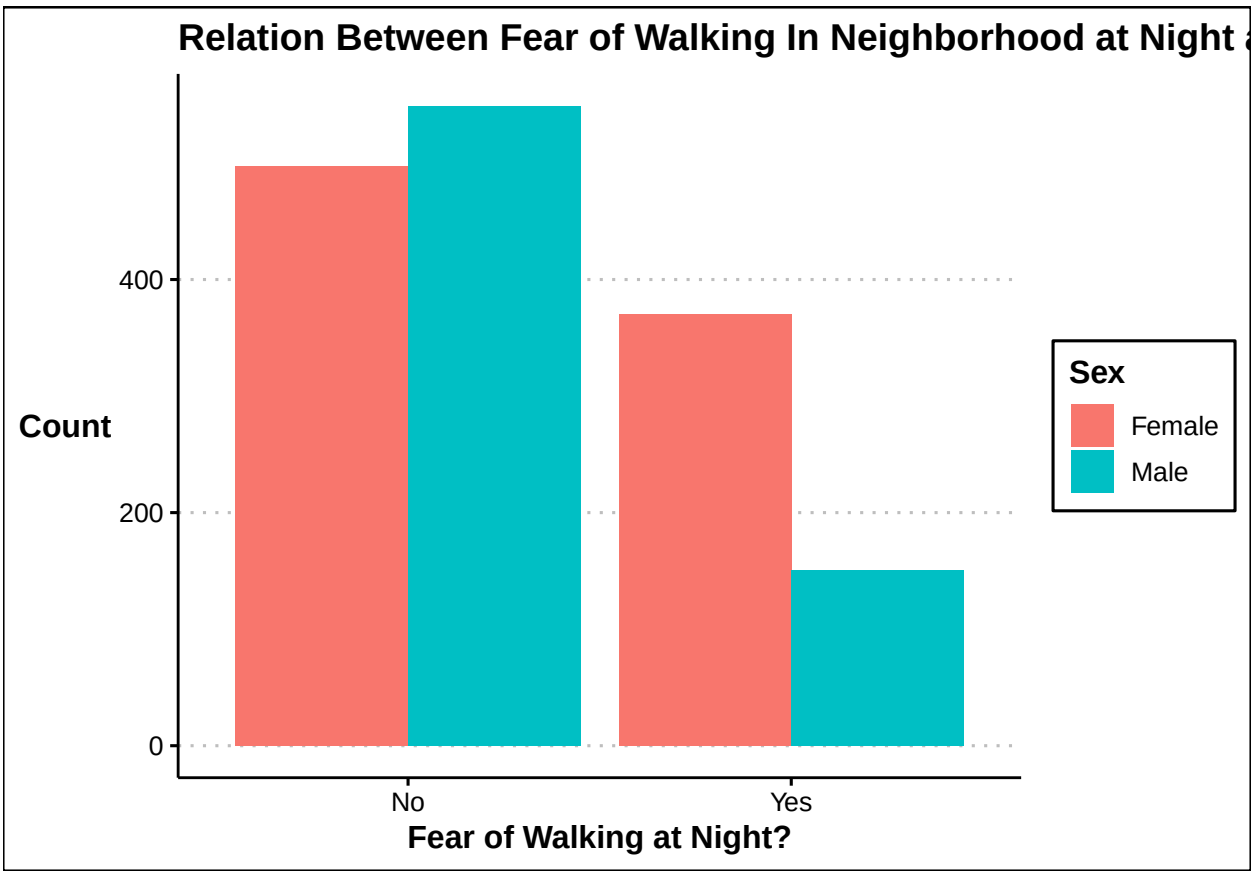
```
## Warning: Removed 20 rows containing non-finite values (stat_boxplot).
```



As we observe this graph, there are very slight differences between the responses at a glance. Both boxes contain the same median. The box that represents those who answered no contains smaller quartiles. What this says is that there are more of those who answered "Yes" than those who answered "No". It also shows that those that answered "Yes" tend to be younger than those who answered "No", therefore, a relation does exist.

Afraid of Walking at Night Alone and Sex

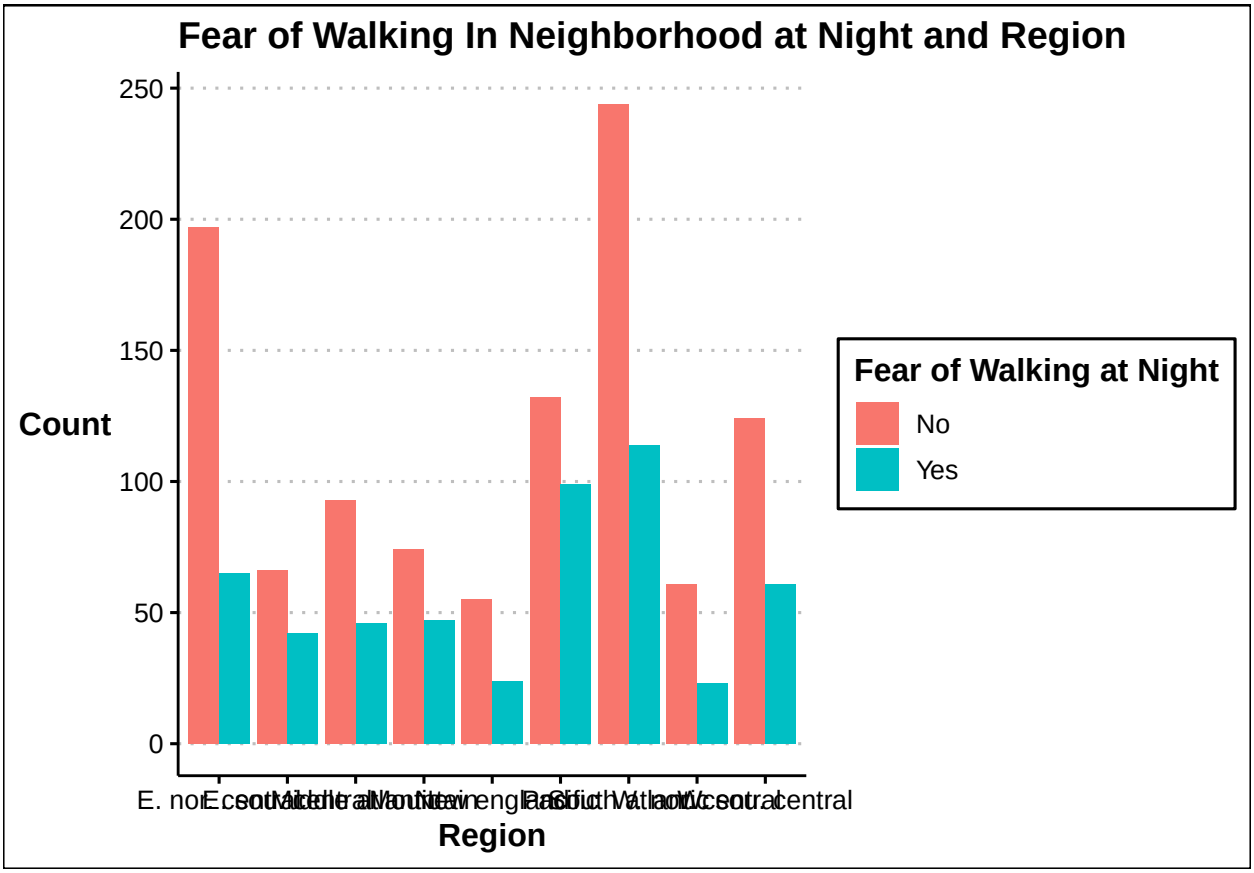
For this question, we took a similar approach to our first graph for this question. We kept the x-axis the same, but changed our fill= to the sex variable instead.



After observing this graph, it is apparent that the majority of recipients answered "No". Within this group, 549 were Male and 497 were Female. For those who answered "Yes", 151 were Male and 370 were Female. We are confident to say that there is a relation that exists here. Females are more likely to fear walking alone at night in their nieghborhoods, whereas Males are not.

Afraid of Walking at Night Alone and Region

The final part of the question pertains to the regions of the country. Therefore, the process of creating the object and graph remains the same, with the only difference being that the x-axis is now region, and our fill = is 'afraidWalkNightNeighborhood'.



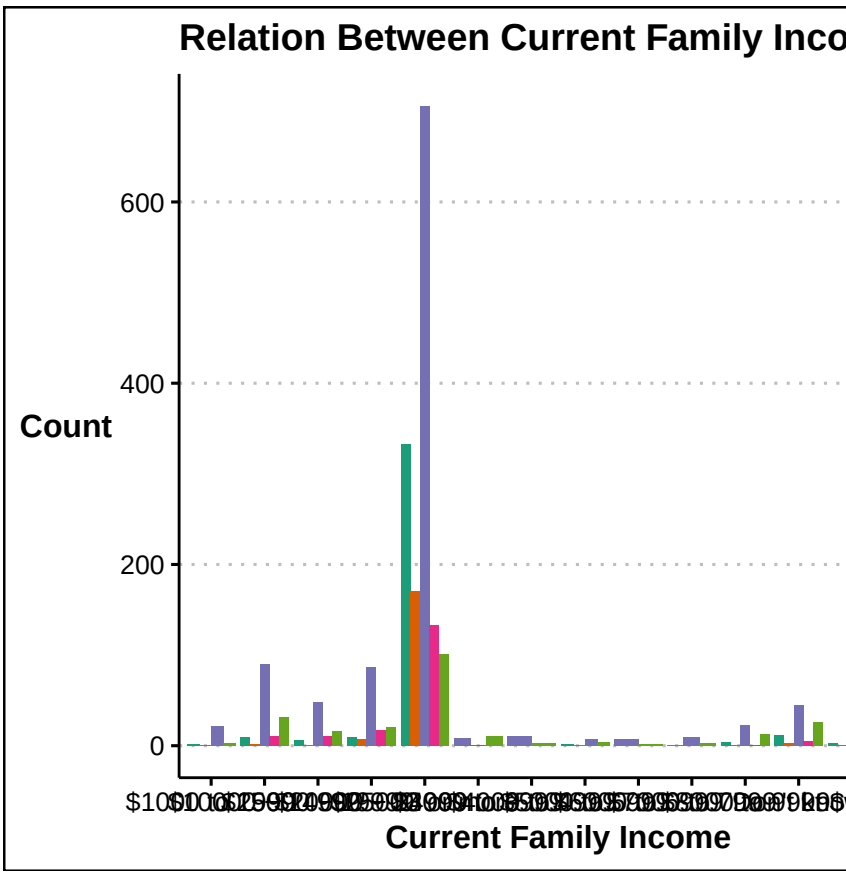
Similarly to our graph for question 2, we see that the South Atlantic region contains the majority of recipients. Across all regions, 1,046 recipents answered "No" and 521 responded "No". It is difficult to say if the region has an affect on walking out at night, therefore we will say that there is no relation between region and a fear of walking out at night. We believe the strongest relation was with the sex of the recipient.

#Question 8: Determining if a relation exists between current family income and highest degree earned

As we analyzed and visualized this sample, we realized that none of the questions asked if there is a relation between current family income and highest degree earned. Revolving Door Consulting felt that it would be particularly helpful to analyze these two variables. We believe this analysis will provide insight for your mission, which is to understand the future of work in the United States.

As before, we created an object, named question8, that uses the 'currentFamIncome' and 'highestDegree' variables.

For our graph, we chose a vertical bar graph, where current family income is on the x-axis, count is on the y-



axis, and highest degree is color coded in the legend.

Similarly to the graph used for the first part of question 7, the majority of recipients live in families than make over $25k. Within this group, the vast majority obtained a High School degree as their highest degree (706). This number alone tops the number of those who acheived a Bachelors (453) and a Graduate Degree (247) across all family income groups. This can mean one of two things; either Bachelor and Gradute Degrees are steadily on the rise for those making 25k or more, or the current labor force does not have a demand for a Bachelor or Graduate degree. If we were able to analyze the same variables across different years, we would be able to have a more definitve answer. Our recommendation to the Foundation is that it explore data collection in this area for future analyses.

Conclusion

After we completed our analysis and visualizations of the sample survey, there were a few conclusions that we can take away.

- The general demographic contains a mix of groups from backgrounds such as: sex, socioeconomic backgrounds, origin, race, and age.
  - The majority demographic are Females who identify as White, within the ages of 30-40, who are currently married
- The representation of race is not fair within this sample since the majority of recipients identify as White
- There exists a relation between hours that someone usually works per week and difficulty sleeping, Stress caused by their jobs, job satisfaction, health, and happiness
- A relation exists between the hours the surveyee and their spouses work per week
- A relation exists one's highest degree earned and the amount of time they spend on their emails

- A relation exists between average hours worked and their household size
- A relation exists between a fear of walking alone in their neighborhood at night and age and sex, but not with current family income and region
- A relation between current family incomes and highest degree earned

We are confident that our findings within the sample survey will guide your organization to new directions of strategy to further your research on the future of work within the United States. Again, I would like to personally thank the Foundation for the Future of Work and Society for choosing Revolving Door Consulting, and I look forward to our future partnerships.