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

```
#loading neccessary packages
library(dplyr)
```

```
##
## 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
```

```
library(readr)
library(ggplot2)
library(viridis)
```

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

## Loading required package: viridisLite
```

```
library(RColorBrewer)
library(ggthemes)
```

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

```
library(ggrepel)
```

```
## 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."

```
#loading dataset
library(readr)
gssProject1 <- read_csv("C:/Users/darie/Downloads/gssProject1.csv")
```

```
## 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.
```

```
View(gssProject1)
```

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.

```
#Getting the final sample size
glimpse(gssProject1)
```

```
## Rows: 2,348
## Columns: 24
## $ X1                       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ...
## $ year                     <dbl> 2018, 2018, 2018, 2018, 2018, 2018, 201...
## $ id                       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ...
## $ race                     <chr> "White", "White", "White", "White", "Bl...
## $ sex                      <chr> "Male", "Female", "Male", "Female", "Ma...
## $ age                      <dbl> 43, 74, 42, 63, 71, 67, 59, 43, 62, 55,...
## $ region                   <chr> "New england", "New england", "New engl...
## $ childhoodFamIncome        <chr> "Below average", "Far below average", "...
## $ highestDegree            <chr> "Junior college", "High school", "Bache...
## $ spouseHighestDegree      <chr> NA, NA, "Junior college", "Junior colle...
## $ householdSize            <dbl> 5, 2, 4, 2, 2, 2, 2, 1, 1, 2, 2, 4, 1, ...
## $ currentFamIncome         <chr> "Refused", "$25000 or more", "$25000 or...
## $ hrsUsuallyWork           <dbl> 41, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ hrsSpouseUsuallyWork     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ maritalStatus            <chr> "Never married", "Separated", "Married"...
## $ numChildren              <dbl> 0, 3, 2, 2, 0, 2, 6, 0, 4, 2, 2, 3, 2, ...
## $ afraidWalkNightNeighborhood <chr> "Yes", "No", NA, NA, "No", "No", "Yes",...
## $ troubleSleeping          <chr> "Sometimes", NA, "Sometimes", "Sometime...
## $ jobSatisfaction          <chr> "Somewhat satisfied", NA, "Somewhat sat...
## $ jobStressful             <chr> "Always", "No issp", "Sometimes", "Ofte...
## $ happiness                <chr> "Pretty happy", "Very happy", "Very hap...
## $ health                   <chr> "Good", "Excellent", NA, NA, "Excellent...
## $ hrsEmailPerWeek          <dbl> 25, NA, 20, 7, NA, 2, NA, NA, 0, 2, 20,...
## $ hrsWWWperWeek            <dbl> 20, NA, 10, 6, NA, 3, NA, NA, 8, 1, 10,...
```

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.

```
#horizontal bar graph to show distribution of race within survey
ggplot(gssProject1, aes(race, fill = race)) +
  geom_bar() +
  scale_fill_manual(values = c("#FE6E02", "#FFCC77", "#152852")) +
  labs(x = "Race",
       y = "Count",
       title = "Distribution of Race in Survey",
       fill = "Race") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
```

```
          axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
          legend.text = element_text(size = 10),
          legend.title = element_text(face="bold", colour="#000000", size=12))
```
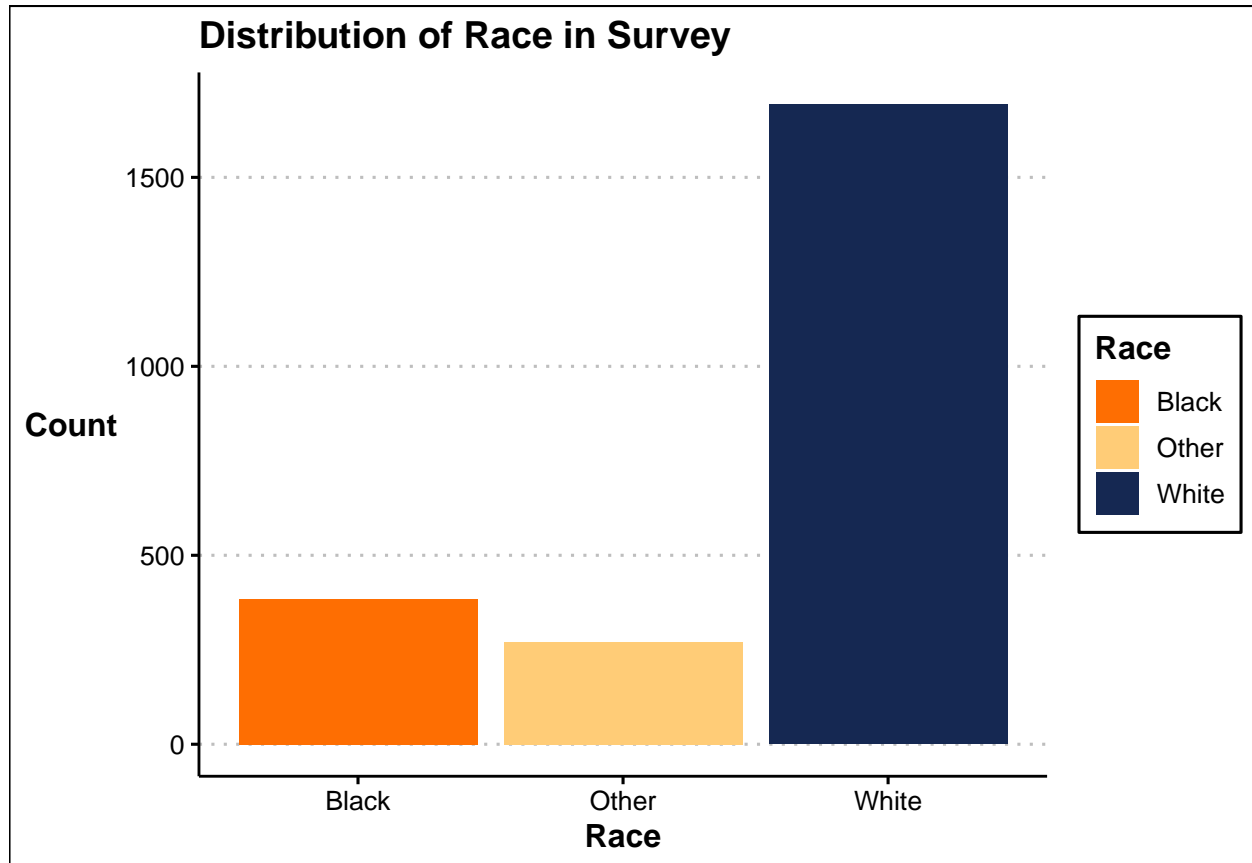


```
#Getting the count of race responses in survey
table(gssProject1$race)
```

```
##
## 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.

```
#Creating a horizontal bar graph to show distribution of sex within survey
ggplot(gssProject1, aes(sex, fill = sex)) +
  geom_bar() +
  scale_fill_manual(values = c("#688052", "#Fa9584")) +
```

```
  labs(x = "Sex",
       y = "Count",
       title = "Distribution of Sex in Survey",
       fill = "Sex") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```
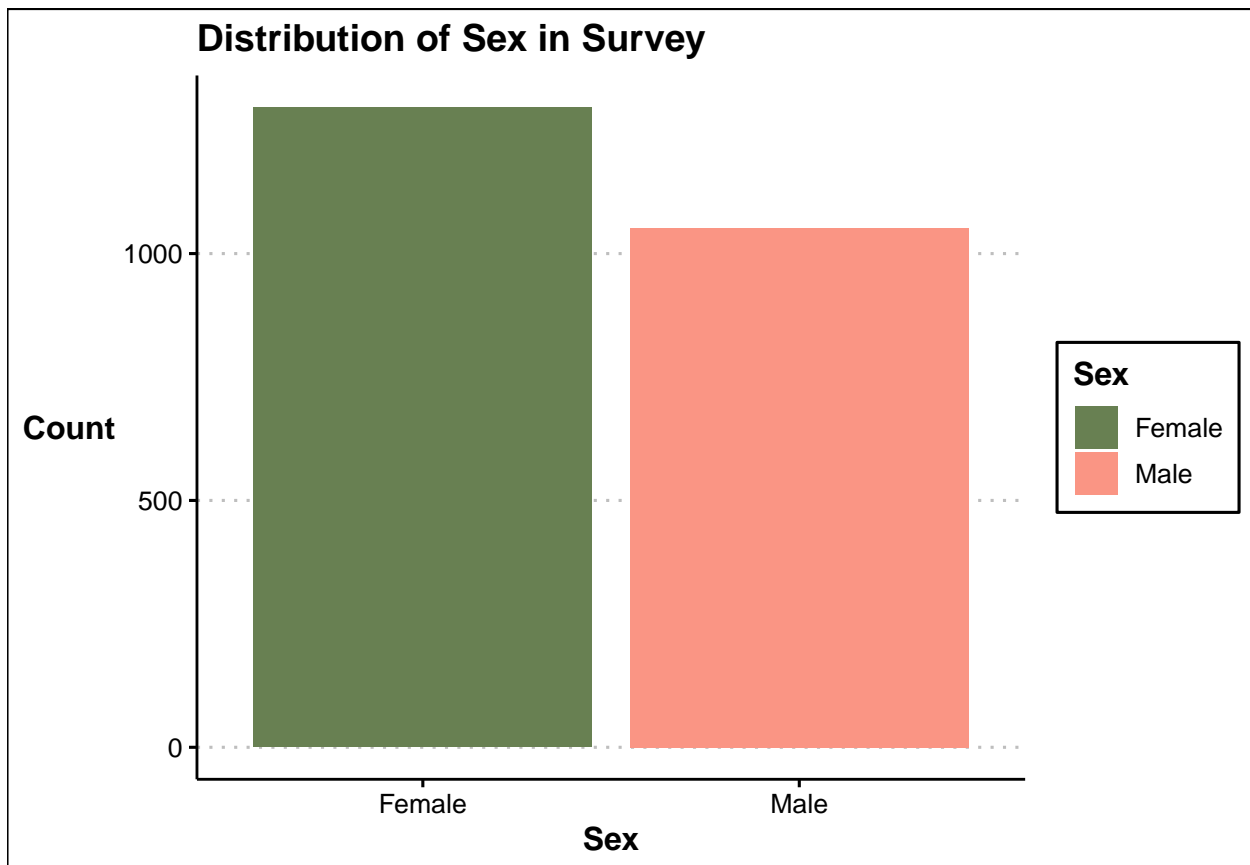


```
#Getting the count of sex responses in survey
table(gssProject1$sex)
```

```
##
## 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.

```
#Creating a horizontal bar graph to show distribution of age within survey
ggplot(gssProject1, aes(age)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  labs(x = "Age",
       y = "Count",
       title = "Distribution of Age in Survey") +
  theme_classic(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5))
```

`## Warning: Removed 36 rows containing non-finite values (stat_bin).`
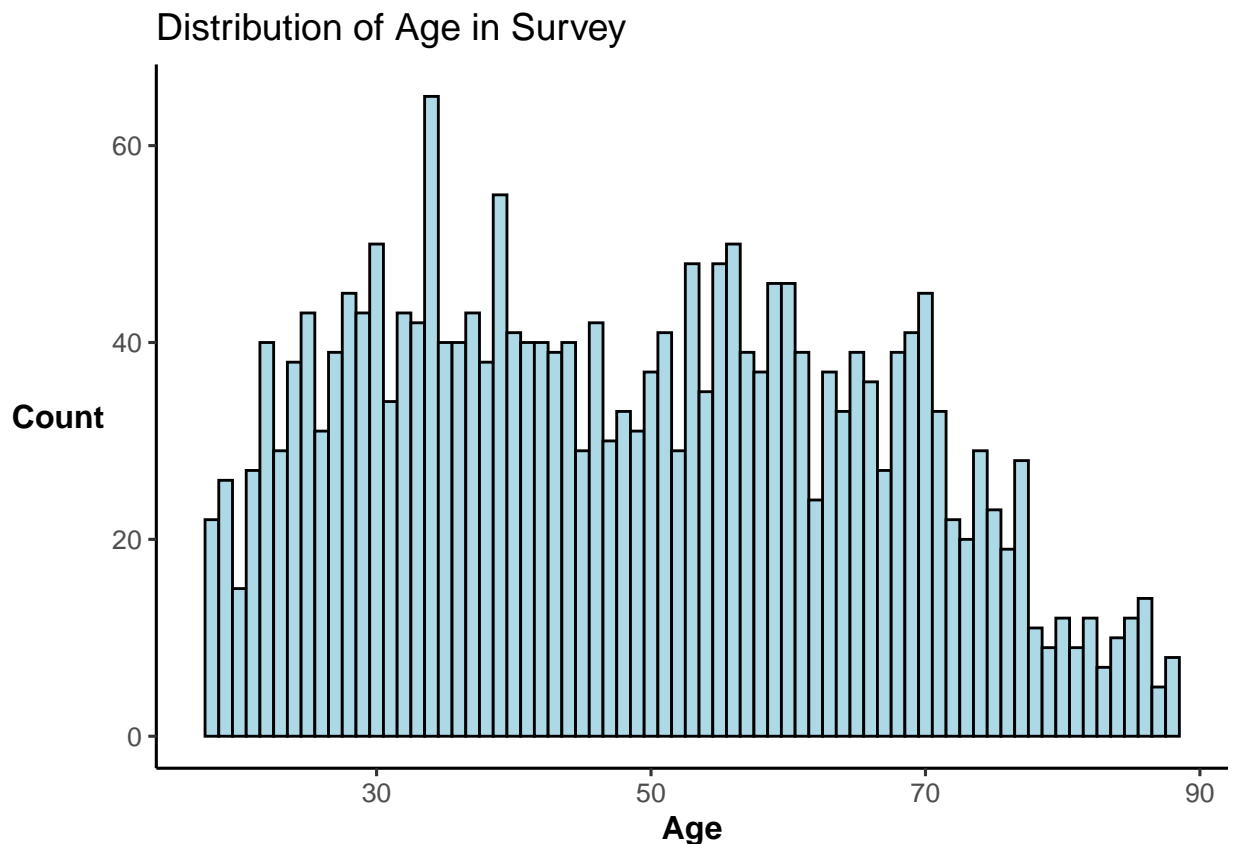
As we can see in this graph, the age groups display a normal distribution. The smallest age group is individuals counted that are 73 years and older. The largest age group within the dataset appears to be individuals between the ages of 30 and 40.

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.

```
#Creating a horizontal bar graph to show distribution of martial status within survey
ggplot(gssProject1, aes(maritalStatus, fill = maritalStatus)) +
  geom_bar() +
  labs(x = "Marital Status",
       y = "Count",
       title = "Distribution of Marital Status in Survey",
       fill = "Martial Status") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```
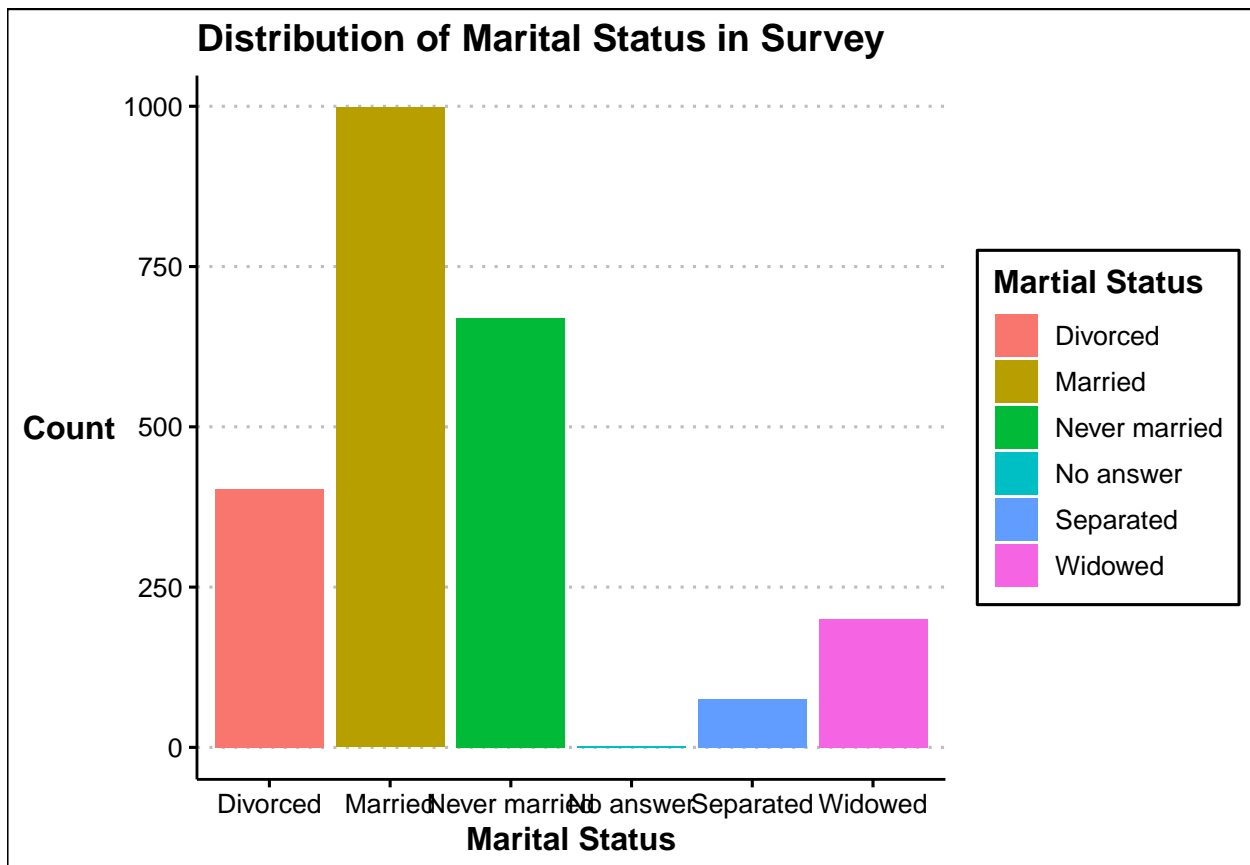


```
#Getting the count of martial status responses in survey
table(gssProject1$maritalStatus)
```

```
##
```

```
##      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.

```
#Creating a stacked bar graph for variables region and race
ggplot(gssProject1, aes(region, fill = race)) +
  geom_bar() +
  labs(x = "Region",
       y = "Count",
       title = "Representation of Race Within Each Region of the Country",
       fill = "Race") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12)) +
  scale_fill_viridis(discrete = T)
```

Representation of Race Within Each Region of the Country
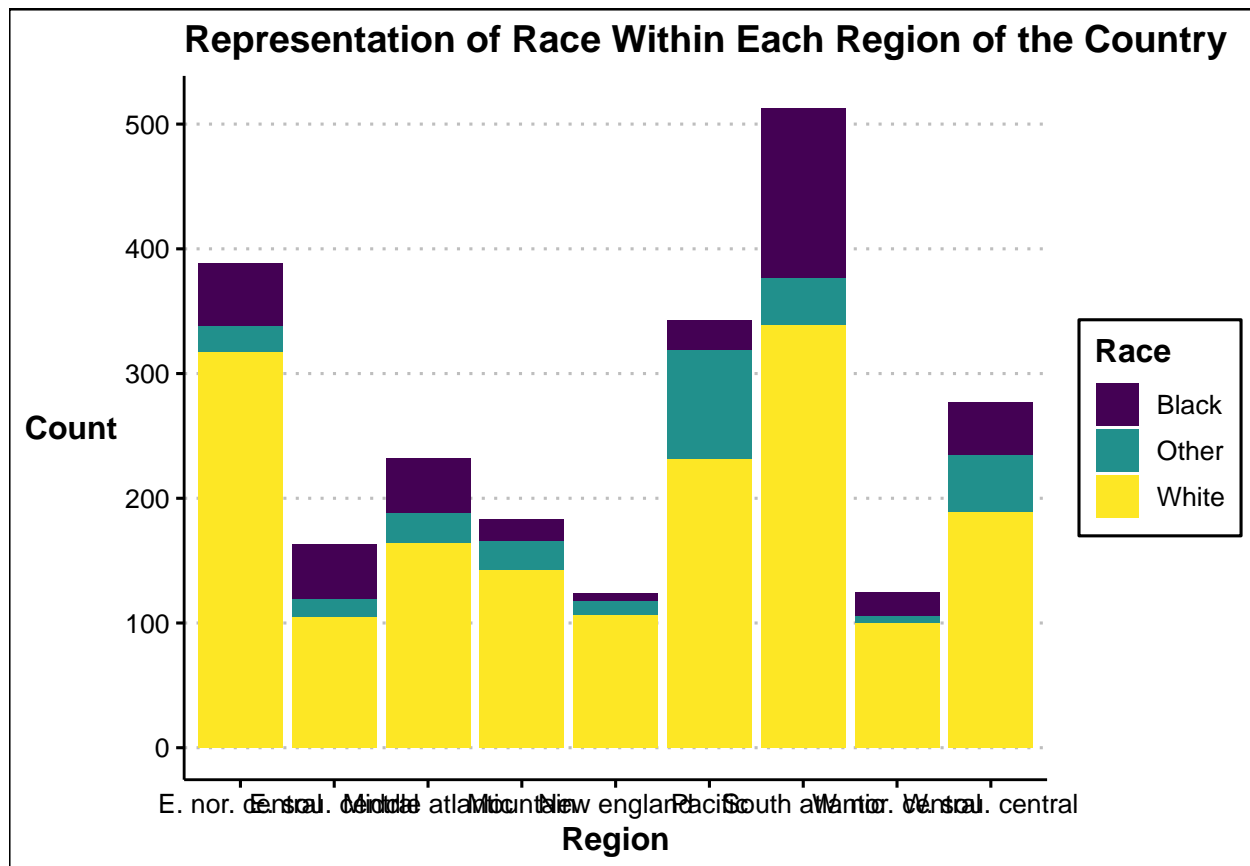

```r
#Creating a table to get counts for the graph
q2_table <- gssProject1 %>%
  select(region, race)

#Display table
table(q2_table)
```

```
##                  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.

```
#Creating new datatset to show the relation between hrsUsuallyWork and troubleSleeping
question3_sleeping <- gssProject1 %>%
  select(hrsUsuallyWork, troubleSleeping) %>%
  filter(!is.na(hrsUsuallyWork)) %>%
  group_by(troubleSleeping) %>%
  summarize(n = n(),
            mean_hrs = mean(hrsUsuallyWork),
            sd = sd(hrsUsuallyWork),
            se = sd/sqrt(n),
            ci = qt(0.975, df = n - 1) * sd / sqrt(n))
```

```
## 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).

```
#Verifying new dataset and getting insight
question3_sleeping
```

```
## # A tibble: 5 x 6
##   troubleSleeping     n mean_hrs    sd    se    ci
## * <chr>           <int>    <dbl> <dbl> <dbl> <dbl>
## 1 Never               2     26    19.8 14.0  178.
## 2 No answer           1     40    NA   NA    NaN
## 3 Often              20     40.2  13.3  2.98   6.23
## 4 Rarely             12     36.2  14.8  4.27   9.39
## 5 Sometimes          13     38.8  15.7  4.34   9.46
```

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.

```r
#plotting vertical graph for hrsUsuallywork and troubleSleeping with error bars
ggplot(question3_sleeping, aes(troubleSleeping, mean_hrs, fill = troubleSleeping)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = mean_hrs - se,
                    ymax = mean_hrs + se),
                width = .1, size = 1) +
  labs(x = "Trouble Sleeping?",
       y = "Average Hours of Work",
       title = "Relation Between Number of Hours of Work and Difficulty of Sleeping",
       fill = "Trouble Sleeping?") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```
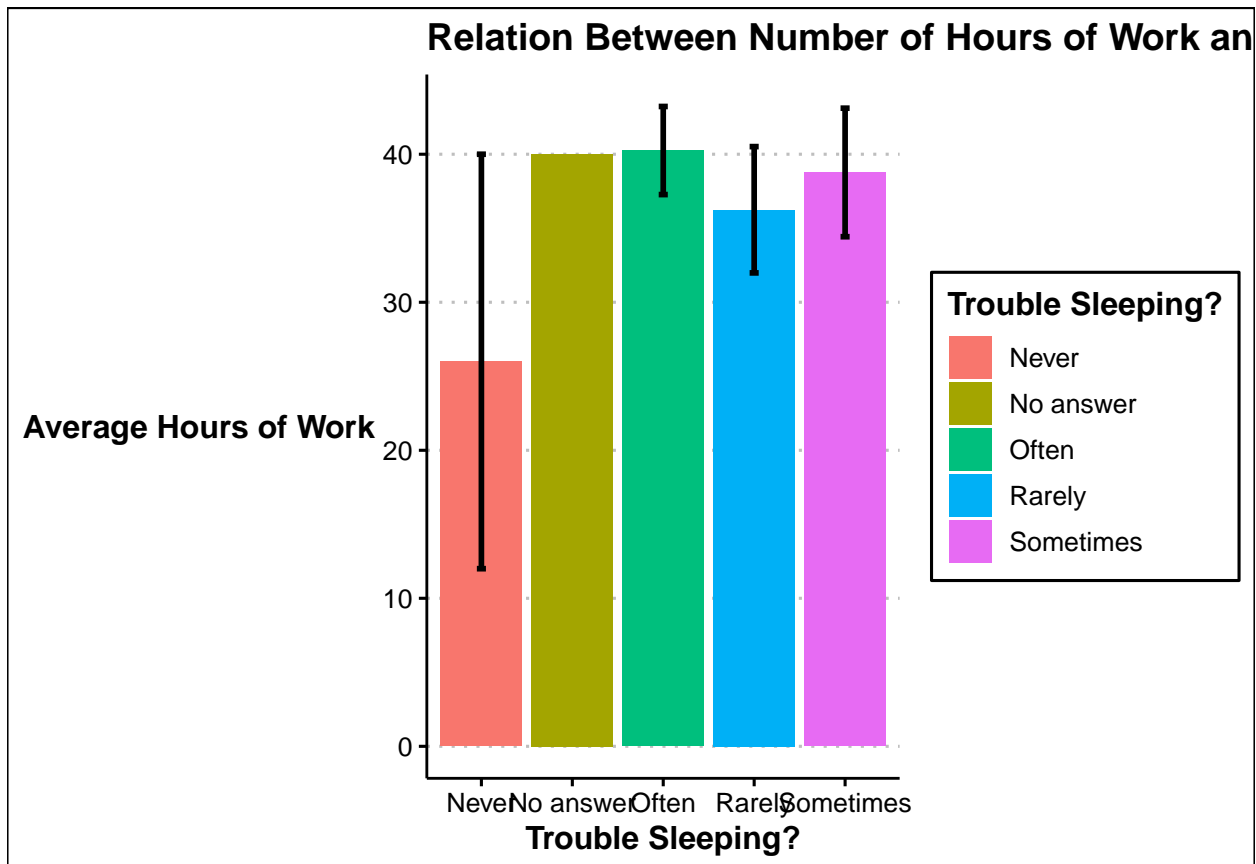

Here, we see that the "Never" response contains the biggest standard error (14.0). The "Often" response yields a standard error of 2.976, which is the lowest of all other responses. This group also works the highest amount of average hours per week (40.25 hours). This gives us the confidence to say that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

```
#Creating new dataset to show the relation between hrsUsuallyWork and jobstress
question3_stress <- gssProject1 %>%
  select(hrsUsuallyWork, jobStressful) %>%
  filter(!is.na(hrsUsuallyWork)) %>%
  group_by(jobStressful) %>%
  summarize(n = n(),
            mean_hrs = mean(hrsUsuallyWork),
            sd = sd(hrsUsuallyWork),
            se = sd/sqrt(n),
            ci = qt(0.975, df = n - 1) * sd / sqrt(n))
```

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

```
#verifying new dataset and getting insight
question3_stress
```

```
## # A tibble: 6 x 6
##   jobStressful     n mean_hrs    sd    se    ci
## * <chr>        <int>    <dbl> <dbl> <dbl> <dbl>
## 1 Always           9     42.4  18.7  6.23  14.4
## 2 Hardly ever      8     31.1  14.5  5.12  12.1
## 3 Never            4     32.5  17.1  8.54  27.2
## 4 No answer        1     40    NA    NA    NaN
## 5 Often           10     39.5  9.85  3.11  7.04
## 6 Sometimes       16     40    13.6  3.39  7.23
```

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.

```
#plotting vertical bargraph with error bars
ggplot(question3_stress, aes(jobStressful, mean_hrs, fill = jobStressful)) +
  geom_bar(stat = "identity") +
    geom_errorbar(aes(ymin = mean_hrs - se,
                      ymax = mean_hrs + se),
                  width = .1, size = 1) +
  labs(x = "Is Your Job Stressful?",
       y = "Average Hours of Work",
       title = "Relation Between Number of Hours of Work and Stress Level",
       fill = "Stress Level") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```

Relation Between Number of Hours of Work an
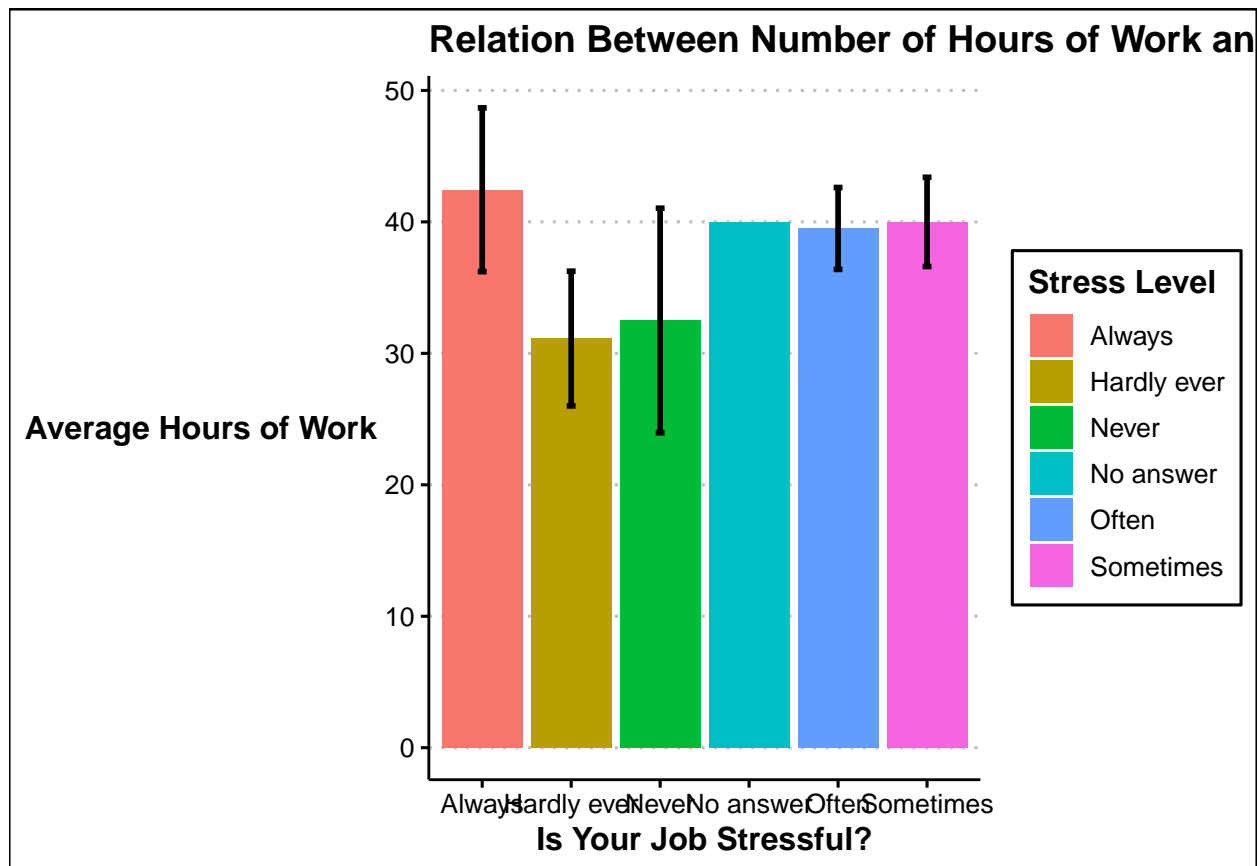
Here, we used the same graphing method as we did for trouble sleeping. The only changes are the x-axis and legend which both contain the jobStressful variable. In this graph, we can see that the standard errors yielded a much lower variance than the trouble sleeping variable. We also see that the group 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.

```r
#Creating new dataset to show the relation between hrsUsuallyWork and jobsatisfaction
question3_jobsatis <- gssProject1 %>%
  select(hrsUsuallyWork, jobSatisfaction) %>%
  filter(!is.na(hrsUsuallyWork)) %>%
  group_by(jobSatisfaction) %>%
  summarize(n = n(),
            mean_hrs = mean(hrsUsuallyWork),
            sd = sd(hrsUsuallyWork),
            se = sd/sqrt(n),
            ci = qt(0.975, df = n - 1) * sd / sqrt(n))
```

```
#verifying new dataset and getting insight
question3_jobsatis
```

```
## # A tibble: 5 x 6
##   jobSatisfaction       n mean_hrs    sd    se    ci
## * <chr>             <int>    <dbl> <dbl> <dbl> <dbl>
## 1 No answer             2     40    0     0     0
## 2 Not at all satisfied  3     36.7  7.64  4.41 19.0
## 3 Not too satisfied     3     34.7  9.24  5.33 22.9
## 4 Somewhat satisfied   21     40.6 11.8   2.57  5.36
## 5 Very satisfied       19     36.3 18.6   4.27  8.96
```

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.

```
#plotting vertical bar graph with error bars
ggplot(question3_jobsatis, aes(jobSatisfaction, mean_hrs, fill = jobSatisfaction)) +
  geom_bar(stat = "identity") +
    geom_errorbar(aes(ymin = mean_hrs - se,
                      ymax = mean_hrs + se),
              width = .1, size = 1) +
  labs(x = "Job Satisfaction",
       y = "Count",
       title = "Relation Between Number of Hours of Work and Job Satisifaction",
       fill = "Job Satisifaction") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```

Relation Between Number of Hours of Work and Job Satisifac
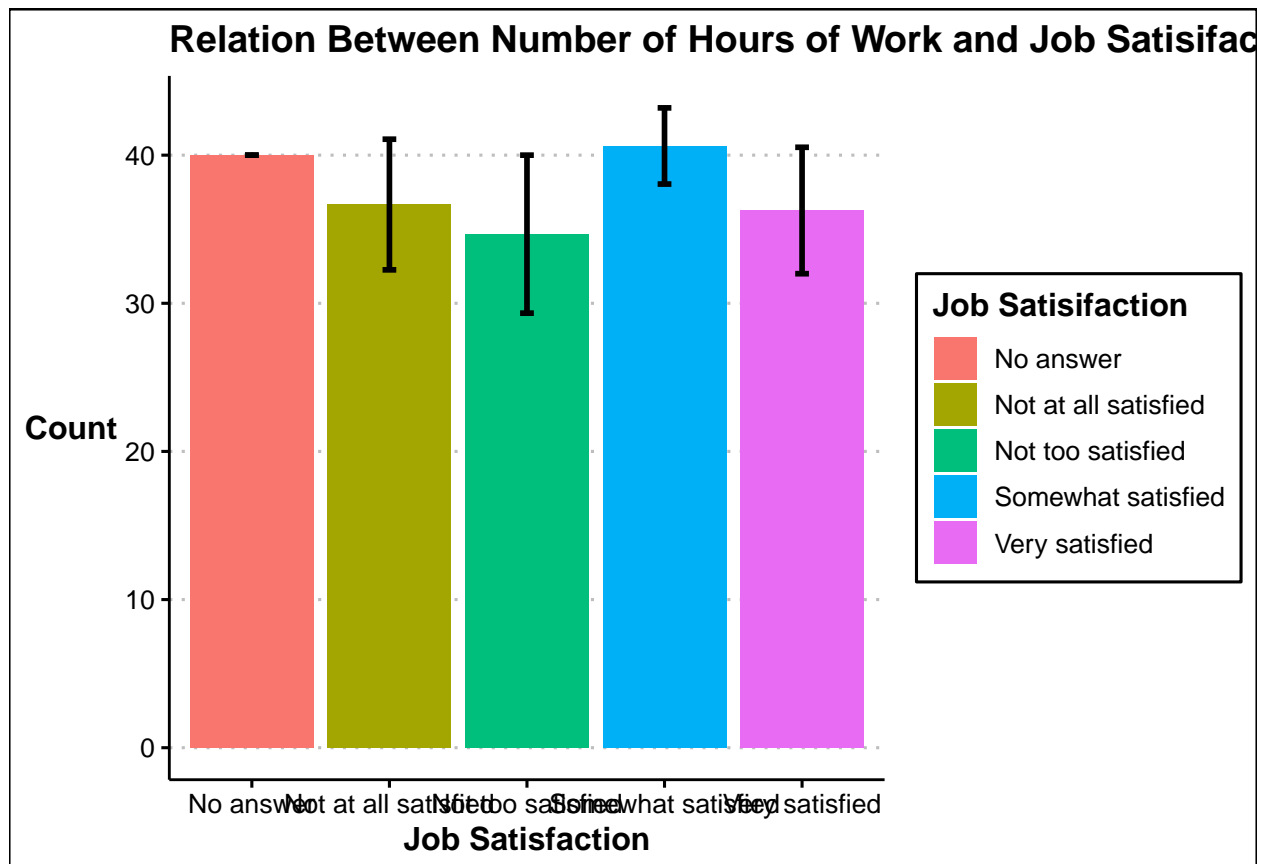
After we analazyed our object and graph, we were able to draw some conclusions. Excluding "No Answer" response, the "Not At All Satisfied" response was lowest (3 recipients). The highest response was "Somewhat Satisfied" (21 recipients). What was suprising is that those who are somewhat satisfied worked 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.

```
#Creating a new dataset to show the relation between hrsUsuallyWork and health
question3_health <- gssProject1 %>%
  select(hrsUsuallyWork, health) %>%
  filter(!is.na(health)) %>%
  filter(!is.na(hrsUsuallyWork))
```

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.

```r
#summarizing new dataset for boxplot
summary(question3_health)
```

```
##   hrsUsuallyWork      health
##   Min.   : 9.00   Length:28
##   1st Qu.:40.00   Class :character
##   Median :40.00   Mode  :character
##   Mean   :39.89
##   3rd Qu.:42.00
##   Max.   :84.00
```

```r
#plotting boxplot
ggplot(question3_health, aes(health, hrsUsuallyWork, fill = health)) +
  geom_boxplot() +
  annotate(geom = "text",
           x = 1,
           y = 84,
           label = "High number of hours worked",
           hjust = 0,
           size = 3) +
  annotate(geom = "text",
           x = 1,
           y = 9,
           label = "Low number of hours worked",
           hjust = 0,
           size = 3) +
  labs(x = "Health",
       y = "Usual Hours Worked",
       title = "Relation Between Number of Hours of Work and Health",
       fill = "Health") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```

Relation Between Number of Hours of Work and Health
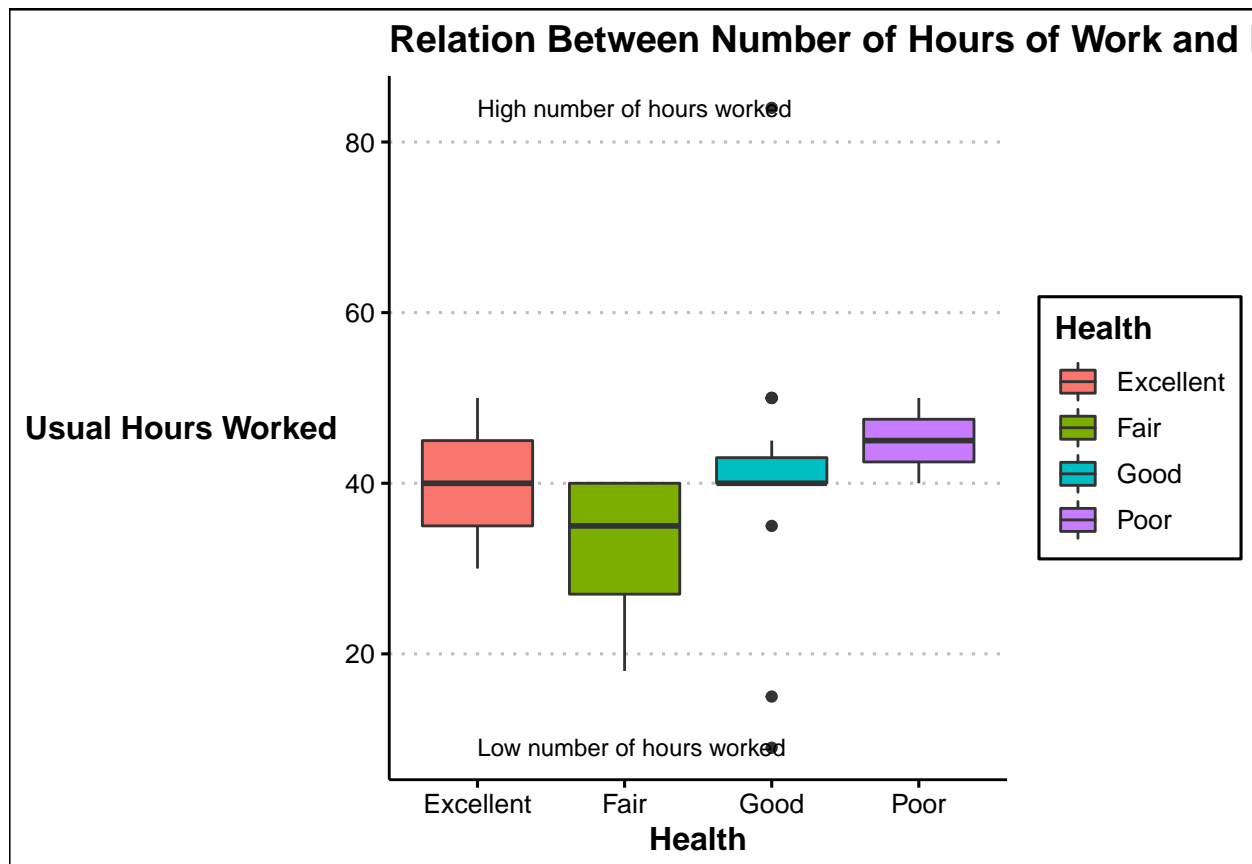
Looking at our summary of the object allows us to better understand the boxplot. Firstly, we will provide an overview of how to interpret a boxplot. The bottom horizontal line in the colored boxes represent the first quartile (the number between the smallest number and the median of the dataset), the 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.

```r
#Creating a new dataset to show the relation between hrsUsuallyWork and Happiness
question3_happy <- gssProject1 %>%
  select(hrsUsuallyWork, happiness) %>%
  filter(!is.na(hrsUsuallyWork))
```

```r
#summarizing new datatset for boxplot
summary(question3_happy)
```

```
##   hrsUsuallyWork    happiness
##   Min.    : 6.00    Length:48
##   1st Qu.:35.00    Class :character
##   Median :40.00    Mode  :character
##   Mean    :38.25
##   3rd Qu.:42.75
##   Max.    :84.00
```

```r
#plotting boxplot
ggplot(question3_happy, aes(happiness, hrsUsuallyWork, fill = happiness)) +
  geom_boxplot() +
  annotate(geom = "text",
           x = 2.05,
           y = 84,
           label = "High number of hours worked",
           hjust = 0,
           size = 3) +
  annotate(geom = "text",
           x = 1.3,
           y = 5,
           label = "Low number of hours worked",
           hjust = 0,
           size = 3) +
  labs(x = "Happiness Level",
       y = "Usual Hours Worked",
       title = "Relation Between Number of Hours of Work and Happiness",
       fill = "Happiness") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```

Relation Between Number of Hours of Work and
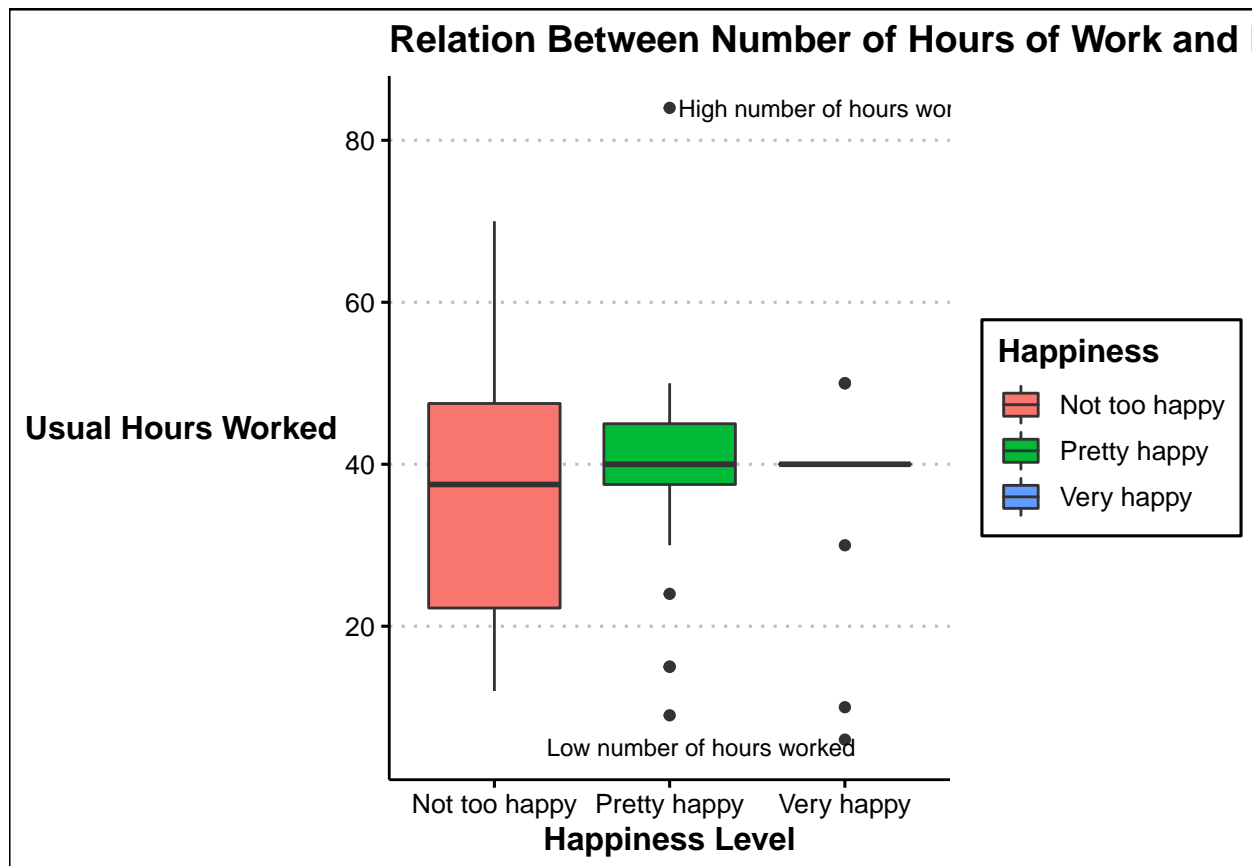
We placed happiness level on x-axis and in the legend.

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'.

```
#Creating a new dataset to show the relation between the numbers of hours spouses usually work each wee

question4 <- gssProject1 %>%
  select(hrsUsuallyWork, hrsSpouseUsuallyWork) %>%
  filter(!is.na(hrsUsuallyWork & hrsSpouseUsuallyWork))

question4
```

```
## # A tibble: 5 x 2
##   hrsUsuallyWork hrsSpouseUsuallyWork
##            <dbl>                <dbl>
## 1             15                   40
## 2             40                   40
## 3             40                   60
## 4             40                   40
## 5             35                   40
```
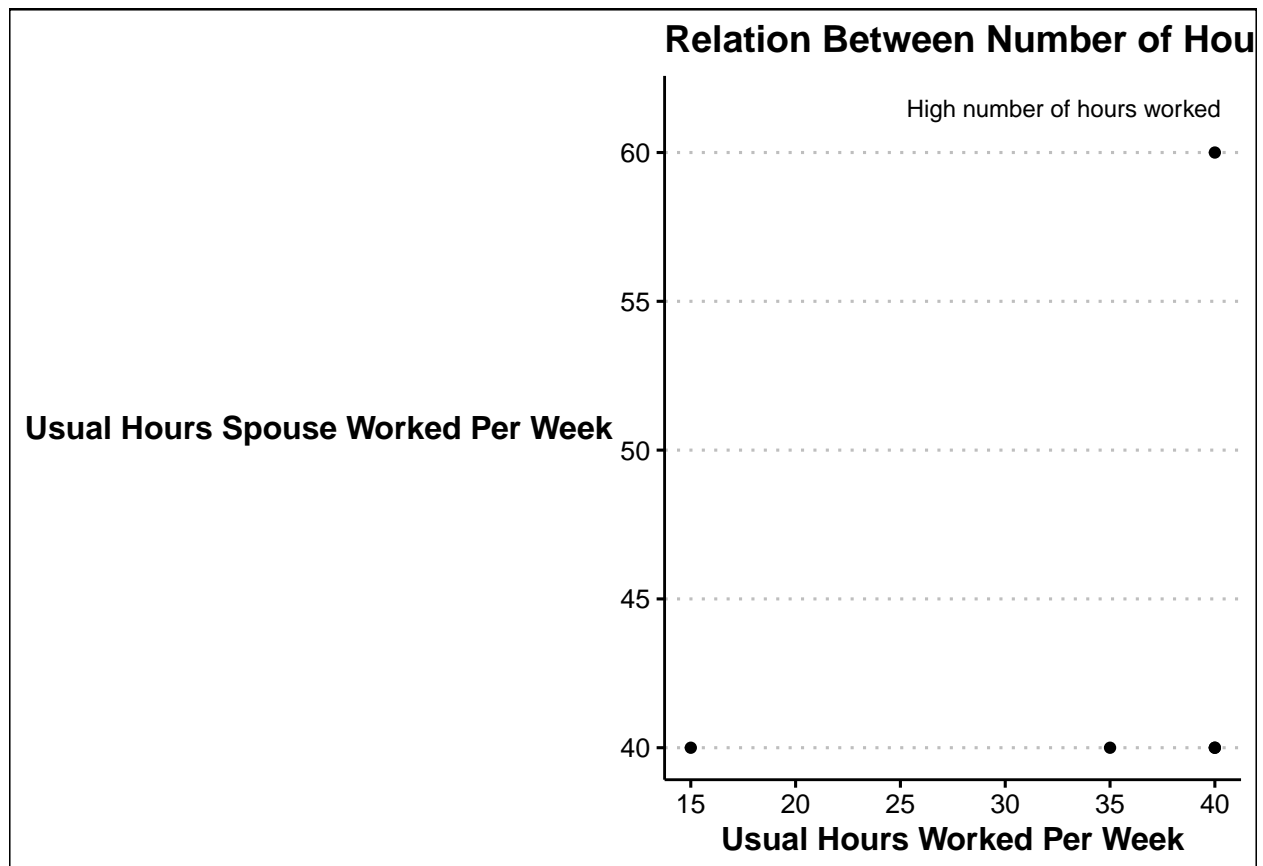
```r
#Getting the count for new dataset
table(question4)
```

```
##                 hrsSpouseUsuallyWork
## hrsUsuallyWork 40 60
##            15  1  0
##            35  1  0
##            40  2  1
```
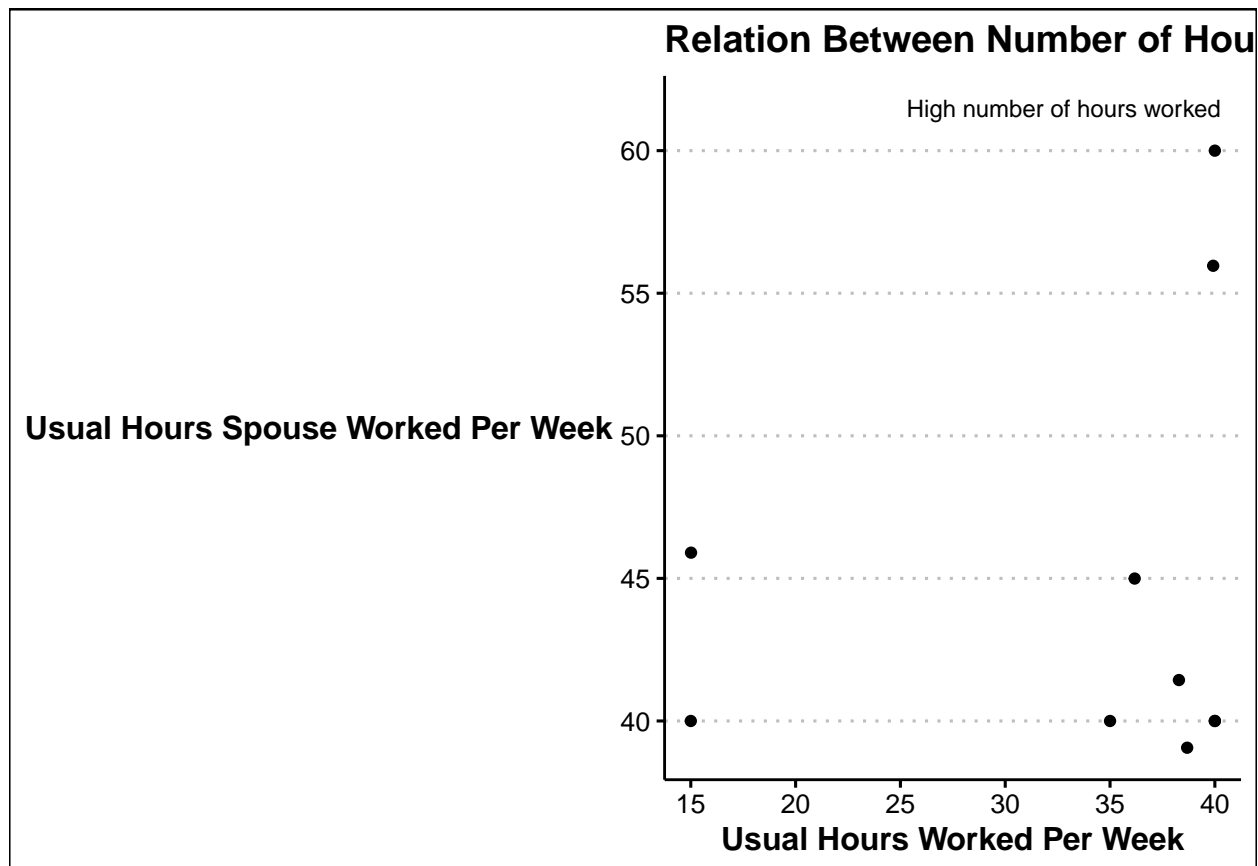
For this object, we can see that hours that the spouse usually works are higher than hours usually worked by the surveyee. It is important to note that, by these value points, we will have overplotted points on our scatterplot. The following graph is a scatterplot wihtout using the geom_jitter() function. 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.

```r
#Plotting scatterplot without geom_jitter for overplotted points
ggplot(question4, aes(hrsUsuallyWork, hrsSpouseUsuallyWork)) +
  geom_point() +
  annotate(geom = "text",
           x = 25.3,
           y = 61.5,
           label = "High number of hours worked",
           hjust = 0,
           size = 3) +
  labs(x = "Usual Hours Worked Per Week",
       y = "Usual Hours Spouse Worked Per Week",
       title = "Relation Between Number of Hours of Work and Number of Hours the Spouse Worked") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```

Relation Between Number of Hou

High number of hours worked

Usual Hours Spouse Worked Per Week

Usual Hours Worked Per Week

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

```
#Plotting scatterplot with geom_jitter for overplotted points
ggplot(question4, aes(hrsUsuallyWork, hrsSpouseUsuallyWork)) +
  geom_point() +
  geom_jitter() +
  annotate(geom = "text",
           x = 25.3,
           y = 61.5,
           label = "High number of hours worked",
           hjust = 0,
           size = 3) +
  labs(x = "Usual Hours Worked Per Week",
       y = "Usual Hours Spouse Worked Per Week",
       title = "Relation Between Number of Hours of Work and Number of Hours the Spouse Worked") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```

Relation Between Number of Hou...

High number of hours worked

Usual Hours Spouse Worked Per Week

Usual Hours Worked Per Week

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.

```
#Creating new dataset to see if people with more education have jobs where they spend more time on emai
question5 <- gssProject1 %>%
  select(hrsEmailPerWeek, highestDegree) %>%
  filter(!is.na(hrsEmailPerWeek)) %>%
  group_by(highestDegree) %>%
  summarize(n = n(),
```

```
            mean_email = mean(hrsEmailPerWeek),
            sd = sd(hrsEmailPerWeek),
            se = sd/sqrt(n),
            ci = qt(0.975, df = n - 1) * sd / sqrt(n))

#verifying new dataset and getting insight
question5
```

```
## # A tibble: 5 x 6
##   highestDegree      n mean_email    sd    se    ci
## * <chr>          <int>      <dbl> <dbl> <dbl> <dbl>
## 1 Bachelor         298      10.3  13.4  0.774 1.52
## 2 Graduate         158      11.3  11.4  0.903 1.78
## 3 High school      719       5.68 11.1  0.412 0.809
## 4 Junior college   121       6.99 12.1  1.10  2.17
## 5 Lt high school   123       3.01  8.17 0.737 1.46
```

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.

```
#Plotting vertical bar graph with errorbars
ggplot(question5, aes(highestDegree, mean_email, fill = highestDegree)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = mean_email - se,
                    ymax = mean_email + se),
                width = .1,
                size = 1) +
  labs(x = "Highest Degree Earned",
       y = "Average Time On Emails",
       title = "Relation Between Average of Hours on Emails and Highest Degree Earned",
       fill = "Highest Degree") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```

Relation Between Average of Hours on Emai
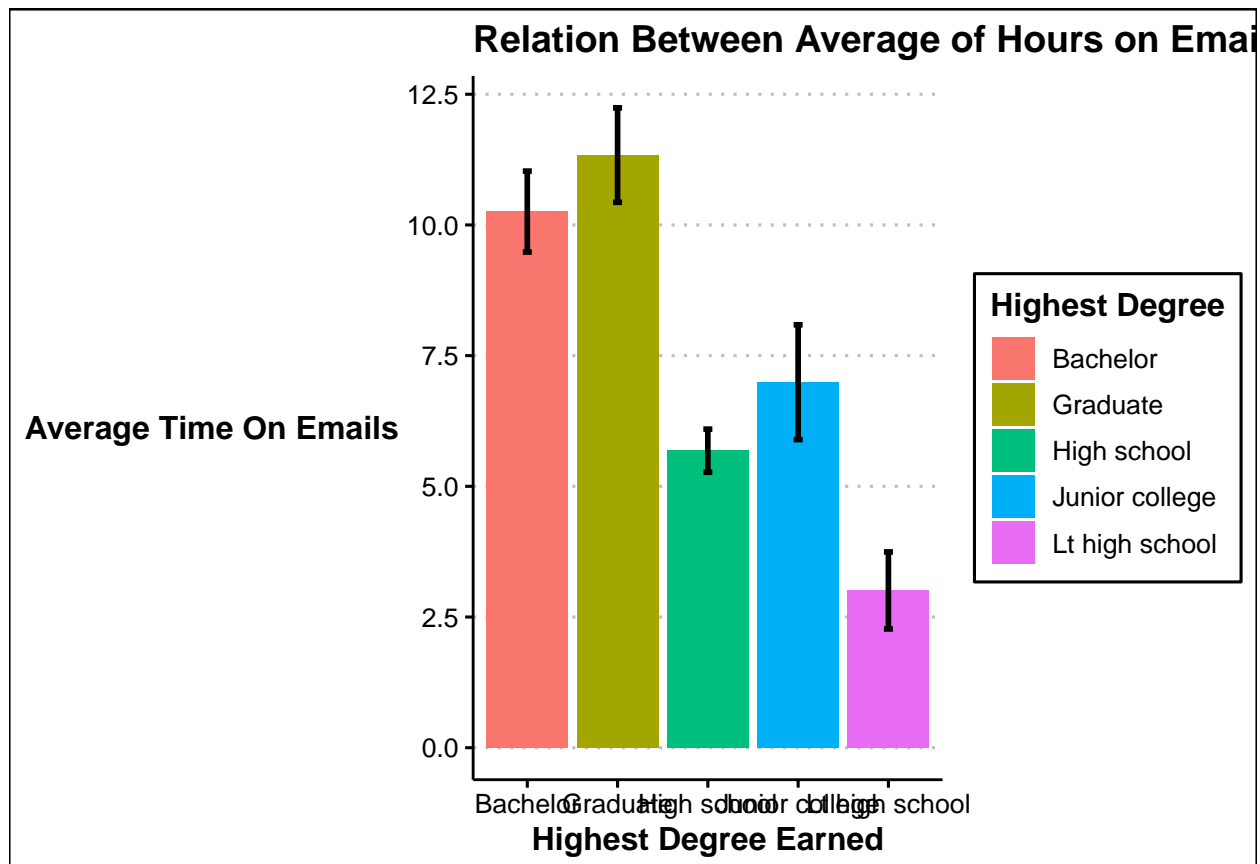
As we observe this visualization, we can confirm that those with a Graduate Degree spend the most time on their emails, while also having the second highest standard error (0.90), which further increases the confidence. Following are Bachelor's degree (10.25 hours), Junior College (6.99 hours), High 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.

```
#Creating a new dataset to figure out if people who live alone work more than people who live with othe
question6 <- gssProject1 %>%
  select(hrsUsuallyWork, householdSize) %>%
  filter(!is.na(hrsUsuallyWork)) %>%
  group_by(householdSize) %>%
  summarize(n = n(),
            mean_hrs = mean(hrsUsuallyWork),
            sd = sd(hrsUsuallyWork),
            se = sd/sqrt(n),
            ci = qt(0.975, df = n - 1) * sd / sqrt(n))
```

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

```
#verifying new dataset and getting insight
question6
```

```
## # A tibble: 6 x 6
##   householdSize     n mean_hrs    sd    se    ci
## *          <dbl> <int>    <dbl> <dbl> <dbl> <dbl>
## 1              1    15     32.3 14.4   3.73  8.00
## 2              2    12     37.9  9.88  2.85  6.27
## 3              3     8     37.8 17.4   6.14 14.5
## 4              4     7     49.1 17.1   6.46 15.8
## 5              5     5     40.2  6.14  2.75  7.62
## 6             11     1     50   NA    NA    NaN
```

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.

```
#Plotting scatterplot with geom_jitter for overplotted points
ggplot(question6, aes(householdSize, mean_hrs)) +
  geom_point() +
  geom_jitter() +
  annotate(geom = "text",
           x = 7,
           y = 51,
           label = "High number of household members",
           hjust = 0,
           size = 3) +
  labs(x = "Household Size",
       y = "Average Hours Worked",
       title = "Relation Between Average of Hours of Work and Household Size") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5))
```

Relation Between Average of Hours of Work an
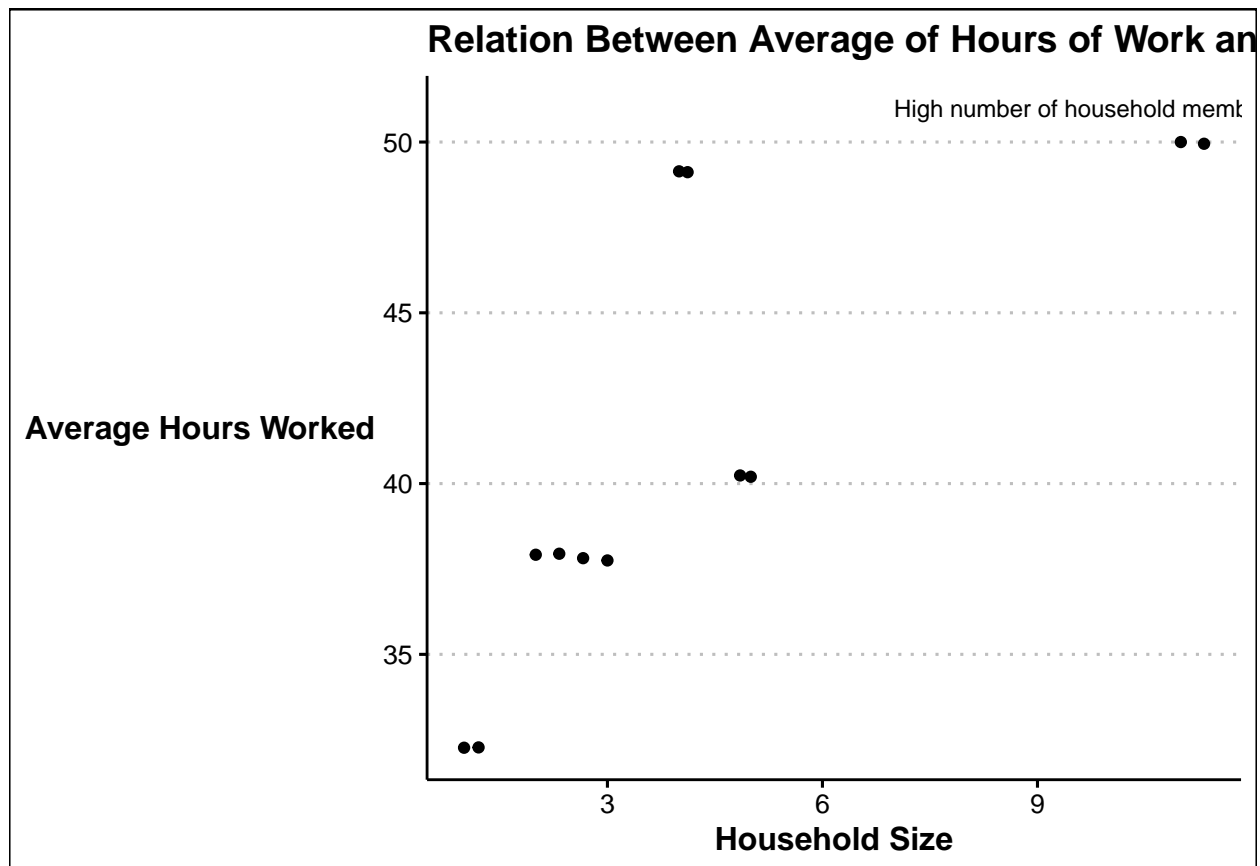
In this graph, household size is on the x-axis, and average hours worked is on the y-axis. For this scatterplot, we used the geom_jitter() function. Here, we observed a clear relation between household size and average hours worked. It is apparent that the bigger the household, the more hours that person 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.

```
#For this question, I am filtering out all of the unneccessary categorical levels except 'Yes' and 'No'
question7 <-  gssProject1 %>%
  filter(afraidWalkNightNeighborhood == "Yes" | afraidWalkNightNeighborhood == "No")

#Verifying filters in new dataset
question7
```

```
## # A tibble: 1,567 x 24
##       X1  year    id race  sex     age region childhoodFamInc~ highestDegree
##    <dbl> <dbl> <dbl> <chr> <chr> <dbl> <chr>  <chr>            <chr>
## 1     1  2018     1 White Male     43 New e~ Below average    Junior colle~
```

```
## 2       2  2018      2 White Fema~     74 New e~ Far below avera~ High school
## 3       5  2018      5 Black Male      71 New e~ Below average     Graduate
## 4       6  2018      6 White Fema~     67 New e~ Average           Bachelor
## 5       7  2018      7 Black Fema~     59 Middl~ Average           High school
## 6       8  2018      8 White Male      43 Middl~ Far below avera~ High school
## 7      11  2018     11 White Male      59 New e~ Above average     Graduate
## 8      13  2018     13 White Fema~     61 New e~ Above average     High school
## 9      14  2018     14 Black Male      44 Middl~ Far below avera~ Bachelor
## 10     16  2018     16 White Fema~     75 Middl~ Average           High school
## # ... with 1,557 more rows, and 15 more variables: spouseHighestDegree <chr>,
## #   householdSize <dbl>, currentFamIncome <chr>, hrsUsuallyWork <dbl>,
## #   hrsSpouseUsuallyWork <dbl>, maritalStatus <chr>, numChildren <dbl>,
## #   afraidWalkNightNeighborhood <chr>, troubleSleeping <chr>,
## #   jobSatisfaction <chr>, jobStressful <chr>, happiness <chr>, health <chr>,
## #   hrsEmailPerWeek <dbl>, hrsWWWperWeek <dbl>
```

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.

```
#Creating new dataset to figure out if there exists a relation between a fear of walking in neighborhoo
question7_income <- question7 %>%
  select(afraidWalkNightNeighborhood, currentFamIncome) %>%
  filter(!is.na(afraidWalkNightNeighborhood)) %>%
  arrange(currentFamIncome)

#Verifying new dataset
question7_income
```

```
## # A tibble: 1,567 x 2
##    afraidWalkNightNeighborhood currentFamIncome
##    <chr>                       <chr>
##  1 Yes                         $1000 to 2999
##  2 Yes                         $1000 to 2999
##  3 Yes                         $1000 to 2999
##  4 No                          $1000 to 2999
##  5 No                          $1000 to 2999
##  6 No                          $1000 to 2999
##  7 Yes                         $1000 to 2999
##  8 Yes                         $1000 to 2999
##  9 No                          $1000 to 2999
## 10 No                          $1000 to 2999
## # ... with 1,557 more rows
```

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

```
#Creating a vertical bar chart
ggplot(question7_income, aes(currentFamIncome, fill = afraidWalkNightNeighborhood)) +
```

```
geom_bar(position = "dodge") +
labs(x = "Current Family Income",
     y = "Count",
     title = "Relation Between Fear of Walking In Neighborhood at Night and Current Family Income",
     fill = "Fear of Walking at Night") +
theme_clean(base_size = 12) +
theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
      axis.text.x  = element_text(size=10),
      axis.text.y = element_text(angle = 0, size = 10),
      axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
      legend.text = element_text(size = 10),
      legend.title = element_text(face="bold", colour="#000000", size=12))
```
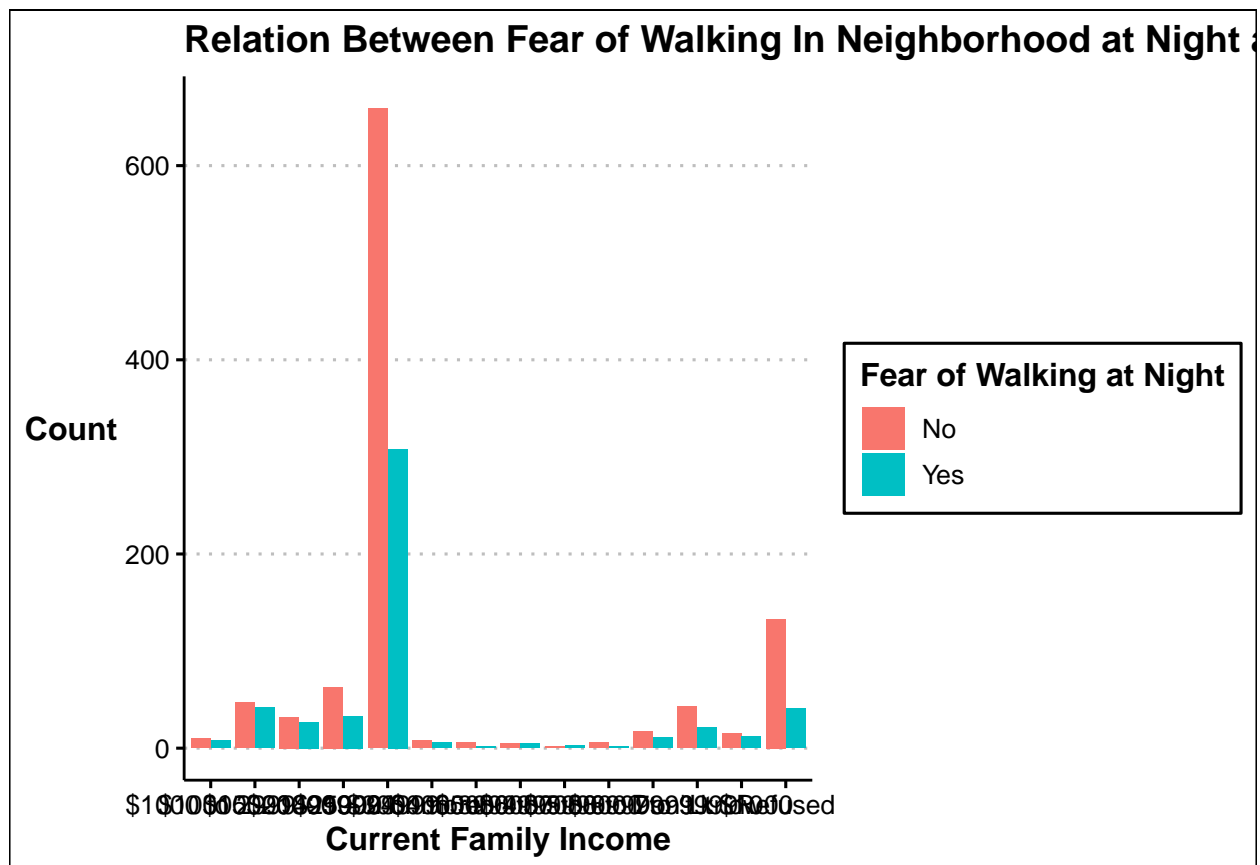

As we observe the graph, it is clear that the majority of recipients live in households that currently make $25k or more per year as a family income; the majority said that they are not afraid to walk at night within their neighborhood. In fact, there is no income level where those who are afriad exceed 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.

```r
#Creating new dataset to figure out of there is a relation between a fear of walking in neighborhood at
question7_age <- question7 %>%
  select(afraidWalkNightNeighborhood, age) %>%
  filter(!is.na(afraidWalkNightNeighborhood)) %>%
  arrange(age)

#Verifying new dataset
question7_age
```

```
## # A tibble: 1,567 x 2
##    afraidWalkNightNeighborhood   age
##    <chr>                       <dbl>
##  1 No                             18
##  2 Yes                            18
##  3 Yes                            18
##  4 No                             18
##  5 No                             18
##  6 No                             18
##  7 No                             18
##  8 No                             18
##  9 Yes                            18
## 10 No                             18
## # ... with 1,557 more rows
```

```r
#Plotting a boxplot
ggplot(question7_age, aes(afraidWalkNightNeighborhood, age, fill = afraidWalkNightNeighborhood )) +
  geom_boxplot() +
  labs(x = "Afraid of Walking at Night",
       y = "Age",
       title = "Fear of Walking In Neighborhood at Night and Age",
       fill = "Fear of Walking at Night") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```

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

Fear of Walking In Neighborhood at Night and Age


```r
#Summarizing new dataset for boxplot
summary(question7_age)
```

```
##  afraidWalkNightNeighborhood      age
##  Length:1567               Min.   :18.00
##  Class :character          1st Qu.:33.00
##  Mode  :character          Median :46.00
##                            Mean   :47.67
##                            3rd Qu.:61.50
##                            Max.   :88.00
##                            NA's   :20
```

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.

```
#Creating a new dataset to figure out of there is a relation between a fear of walking in neighborhood
question7_sex <- question7 %>%
  select(afraidWalkNightNeighborhood, sex) %>%
  filter(!is.na(afraidWalkNightNeighborhood))

#Verifying new dataset
question7_sex
```

```
## # A tibble: 1,567 x 2
##    afraidWalkNightNeighborhood sex
##    <chr>                       <chr>
##  1 Yes                         Male
##  2 No                          Female
##  3 No                          Male
##  4 No                          Female
##  5 Yes                         Female
##  6 No                          Male
##  7 Yes                         Male
##  8 Yes                         Female
##  9 Yes                         Male
## 10 Yes                         Female
## # ... with 1,557 more rows
```

```
#Creating a vertical bar chart
ggplot(question7_sex, aes(afraidWalkNightNeighborhood, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(x = "Fear of Walking at Night?",
       y = "Count",
       title = "Relation Between Fear of Walking In Neighborhood at Night and Sex",
       fill = "Sex") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```
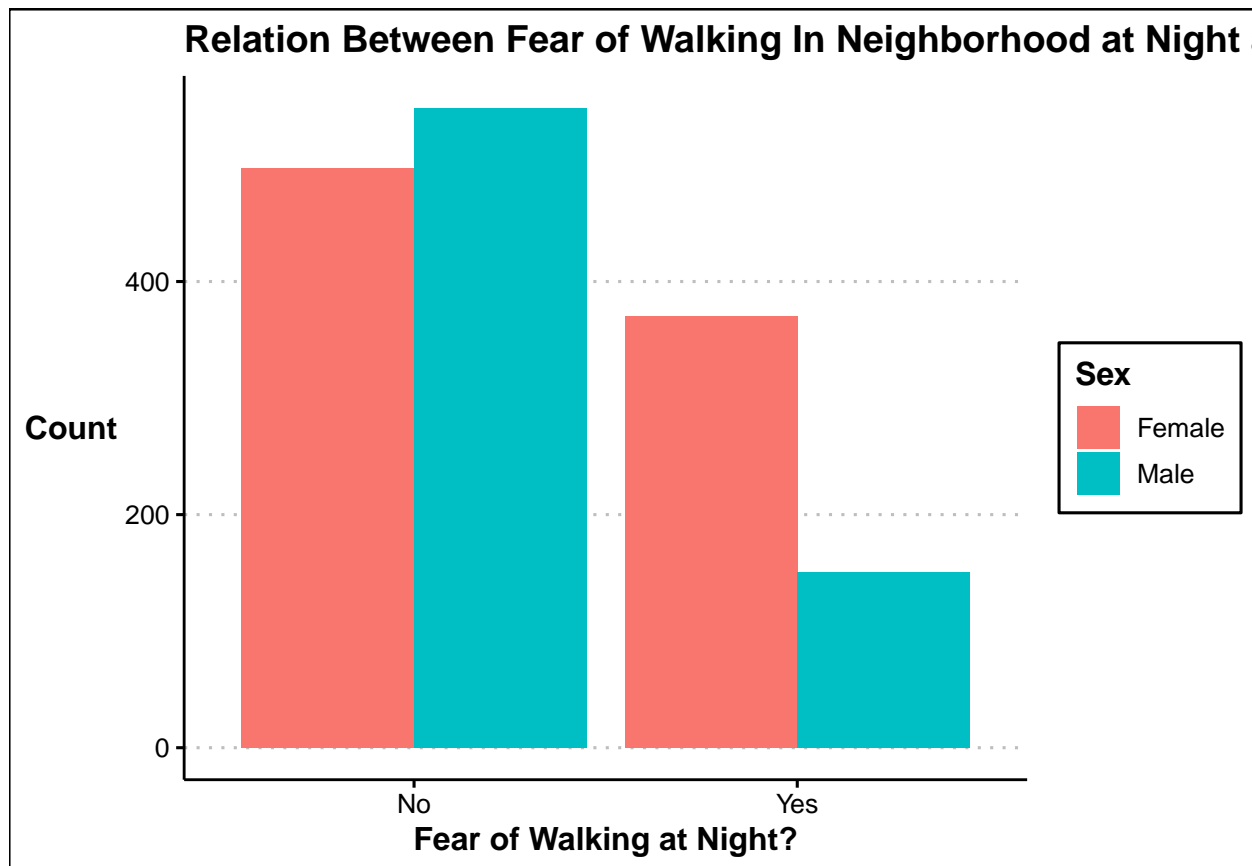
```
#Using a table to get a count for this dataset
table(question7_sex)
```

```
##                        sex
## afraidWalkNightNeighborhood Female Male
##                         No    497  549
##                         Yes   370  151
```

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'.

```
#Creating a dataset to figure out if there is a relation between a fear of walking in neighborhood at n
question7_region <- question7 %>%
  select(afraidWalkNightNeighborhood, region) %>%
  filter(!is.na(afraidWalkNightNeighborhood))
```

```
#Verifying new dataset
question7_region
```

```
## # A tibble: 1,567 x 2
##    afraidWalkNightNeighborhood region
##    <chr>                       <chr>
##  1 Yes                         New england
##  2 No                          New england
##  3 No                          New england
##  4 No                          New england
##  5 Yes                         Middle atlantic
##  6 No                          Middle atlantic
##  7 Yes                         New england
##  8 Yes                         New england
##  9 Yes                         Middle atlantic
## 10 Yes                         Middle atlantic
## # ... with 1,557 more rows
```

```
#Creating stacked bar chart
ggplot(question7_region, aes(region, fill = afraidWalkNightNeighborhood)) +
  geom_bar(position = "dodge") +
  labs(x = "Region",
       y = "Count",
       title = "Fear of Walking In Neighborhood at Night and Region",
       fill = "Fear of Walking at Night") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12))
```

Fear of Walking In Neighborhood at Night and Region
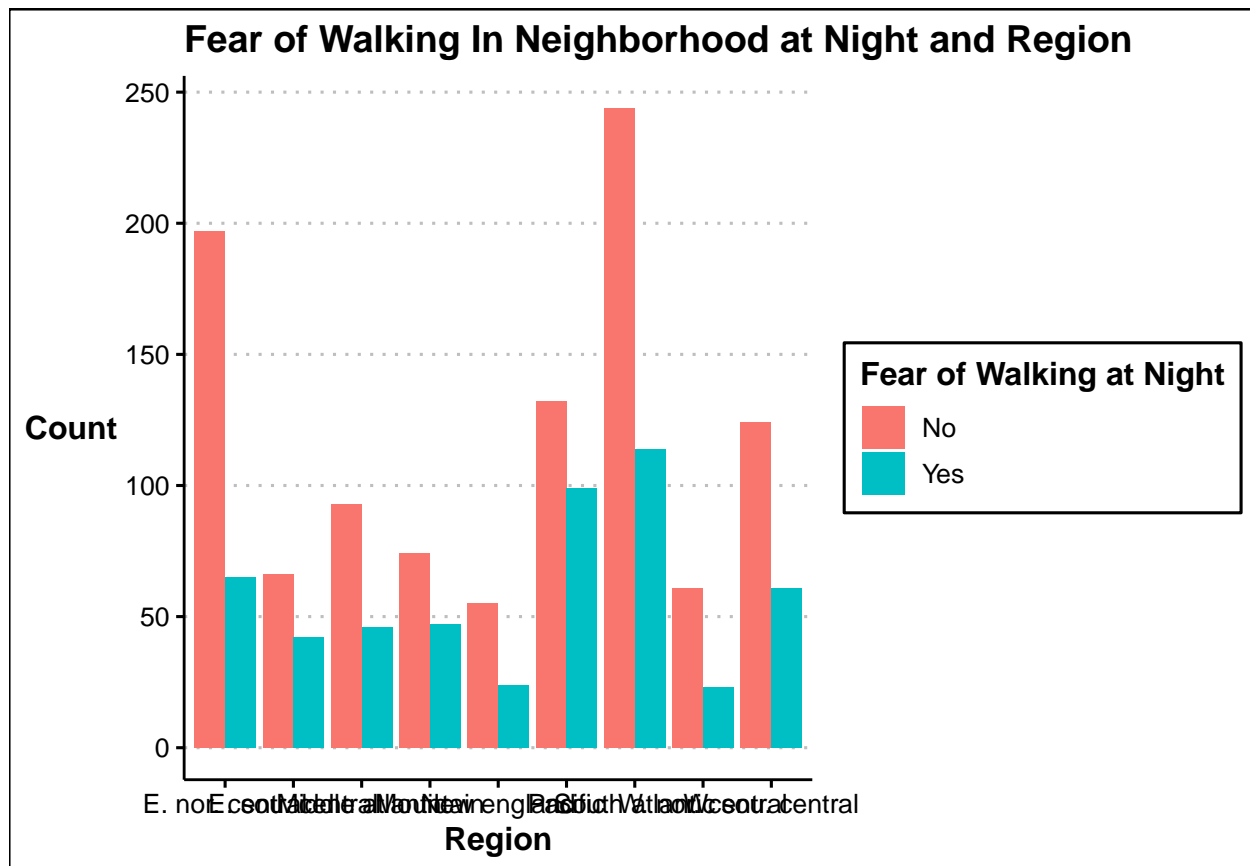
```
#Using a table to get the counts for the new dataset
table(question7_region)
```

```
##                                 region
## afraidWalkNightNeighborhood E. nor. central E. sou. central Middle atlantic
##                         No               197              66              93
##                         Yes               65              42              46
##                                 region
## afraidWalkNightNeighborhood Mountain New england Pacific South atlantic
##                         No         74          55     132           244
##                         Yes        47          24      99           114
##                                 region
## afraidWalkNightNeighborhood W. nor. central W. sou. central
##                         No               61             124
##                         Yes              23              61
```

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.

```
#creating a new dataset to compare current family income to their highest degree earned. This can show
question8 <- gssProject1 %>%
  select(currentFamIncome, highestDegree)

#Verifying new dataset
question8
```

```
## # A tibble: 2,348 x 2
##    currentFamIncome highestDegree
##    <chr>            <chr>
##  1 Refused          Junior college
##  2 $25000 or more   High school
##  3 $25000 or more   Bachelor
##  4 Refused          Bachelor
##  5 Refused          Graduate
##  6 Don't know       Bachelor
##  7 $15000 - 19999   High school
##  8 $25000 or more   High school
##  9 $5000 to 5999    Lt high school
## 10 $25000 or more   High school
## # ... with 2,338 more rows
```

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.

```
#Plotting vertical bar graph
ggplot(question8, aes(currentFamIncome, fill = highestDegree)) +
  geom_bar(position = "dodge") +
  labs(x = "Current Family Income",
       y = "Count",
       title = "Relation Between Current Family Income and Highest Degree Earned",
       fill = "Highest Degree") +
  theme_clean(base_size = 12) +
  theme(axis.title.x = element_text(face="bold", colour="#000000", size=12),
        axis.text.x  = element_text(size=10),
        axis.text.y = element_text(angle = 0, size = 10),
        axis.title.y = element_text(angle=0, size = 12, face = "bold", vjust = .5),
        legend.text = element_text(size = 10),
        legend.title = element_text(face="bold", colour="#000000", size=12)) +
  scale_fill_brewer(palette = "Dark2")
```

Relation Between Current Family Income and Highest Degree
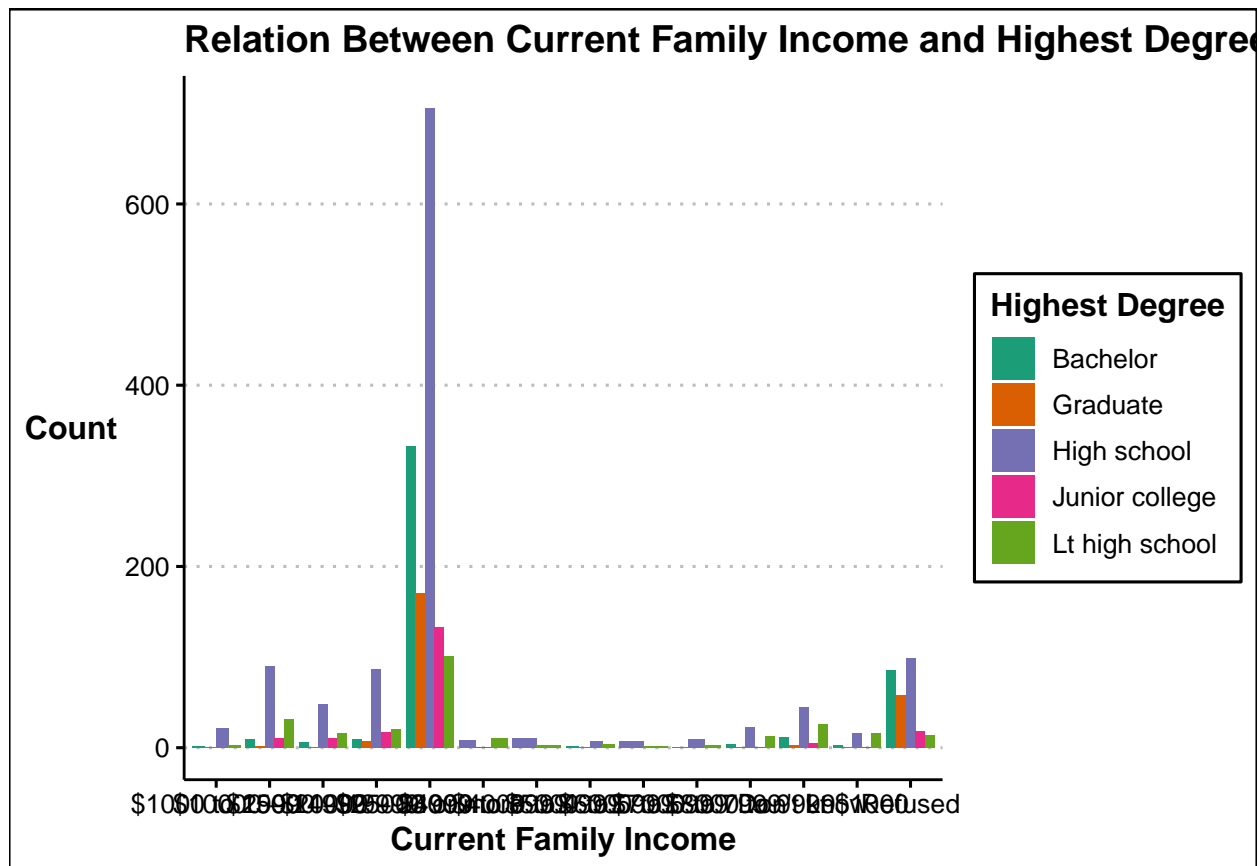
```
#Using a table to get counts for new dataset
table(question8)
```

```
##                    highestDegree
## currentFamIncome Bachelor Graduate High school Junior college Lt high school
##     $1000 to 2999        2        1         22              0              3
##     $10000 - 14999       9        2         90             10             31
##     $15000 - 19999       6        1         48             10             16
##     $20000 - 24999       9        7         87             17             20
##     $25000 or more     333      171        706            133            101
##     $3000 to 3999        0        0          8              1             10
##     $4000 to 4999        0        0         11              0              3
##     $5000 to 5999        2        1          7              0              4
##     $6000 to 6999        0        0          7              0              2
##     $7000 to 7999        0        1          9              0              3
##     $8000 to 9999        4        1         23              1             13
##     Don't know          12        3         45              5             26
##     Lt $1000             3        1         16              1             16
##     Refused             85       58         99             18             14
```

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

36

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.