m158-project1-NHANES

NHANES Data Description:

We are using data from the National Health and Nutrition Exam Survey (NHANES) database that was collected between 2009-2012 with adjusted weighing.

The target population of NHANES is "the non-institutionalized civilian resident population of the United States". Therefore, the observational unit is a civilian resident of the United States (of any age).

There are 75 total variables in NHANES data. However, only 10 are relevant to our project.

Relevant Variables:

Age (quantitative) - Age in years at screening of study participant. All subjects 80 and older were recorded as 80.

Testosterone (quantitative) - Testerone total (ng/dL), recorded for patients 6 and older. Note that no testosterone data for 2009-2010 was recorded.

Physactivedays (numerical discrete) - Number of days in a typical week that partipant does moderate or vigorous-intensity activity. Reported in patients 12 years or older.

Pulse (quantitative) - 60 second pulse rate

HHIncomeMid (numerical) - Total annual gross income for the household in US dollars, derived from the median of each partition. Variable was partitioned into blocks with the smallest one being (0,4999) and the largest block being (100,000 or more).

Gender (categorical) - Gender (sex) of study participant coded as "male" or "female".

SleepTrouble (categorical) - Paritipant has told a doctor or other health professional that they had trouble sleeping. Reported in patients 16 and older. Either "yes" or "no".

Depressed (categorical) - Self reported number of days where participant felt down, depressed or hopeless. Reported in patients 18 and olders with categories of "none", "several", "majority (more than half the days)", or "almostall".

Education (categorical) - Educational level of study participant. Reported for ages 20 or older. Categories to choose from are "8thgrade", "9-11thgrade", "Highschool", "SomeCollege", or "CollegeGrad".

Work (categorical) - Categorizes whether study participant is "working", "not working" or no data was collected.

Filtering NHANES for to contain only our relevant variables:

```
OurData <- NHANES %>% select(Age, HHIncomeMid, Testosterone, PhysActiveDays,
                             Pulse, Gender, SleepTrouble, Depressed,
                             Education, Work)
names(OurData)
```

```
##  [1] "Age"            "HHIncomeMid"    "Testosterone"   "PhysActiveDays"
##  [5] "Pulse"          "Gender"         "SleepTrouble"   "Depressed"
##  [9] "Education"      "Work"
```

```
dim(OurData)
```

```
## [1] 10000    10
```

Summary Statistics

```
skim(OurData) %>%
  dplyr::select(skim_variable, numeric.mean, numeric.sd, complete_rate, numeric.p50)
```

```
## # A tibble: 10 x 5
##    skim_variable  numeric.mean numeric.sd complete_rate numeric.p50
##    <chr>                 <dbl>      <dbl>         <dbl>       <dbl>
##  1 Gender                NA         NA            1            NA
##  2 SleepTrouble          NA         NA            0.777        NA
##  3 Depressed             NA         NA            0.667        NA
##  4 Education             NA         NA            0.722        NA
##  5 Work                  NA         NA            0.777        NA
##  6 Age                   36.7       22.4          1            36
##  7 HHIncomeMid        57206.     33020.           0.919     50000
##  8 Testosterone         198.       227.           0.413        43.8
##  9 PhysActiveDays         3.74       1.84         0.466         3
## 10 Pulse                 73.6       12.2          0.856        72
```

We can see that our data has varying rates of completion with most being over 50% reported, but Testosterone and PhysActiveDays have 41.25% and 46.6% completion rates respectively. As such, it is likely that our statistics for Testorsterone and PhysActievDays are not fully representative of the population.

Additionally, Testosterone has a extremely high standard deviation (SD) relative to its mean, indicating the data distribution may be skewed. This is corroborated by the mean and median (p.50) being substantially different for Testosterone, and so we expect the distribution to be skew right.

For all the other variables, median (p.50) and mean seems relatively close indicating the data distribution is somewhat symmetrical. However, to confirm this fact, we will have to look at histograms and bar graphs to see if there is a skewness to the data distribution.

Each standard deviation (SD) also tells us a little more information about the shape of the data. For example, for age the SD is relatively high but the measures of center are about the same. So, we would expect age to have a more uniform distribution. HHincomeMid seems to have a high SD relative to its center, so we would expect a slightly skewed data distribution even though the mean and median seem relatively close.

Graphs

Skew right,_____...

```
OurData %>%
  ggplot(aes(x = Age)) +
  geom_histogram()
```

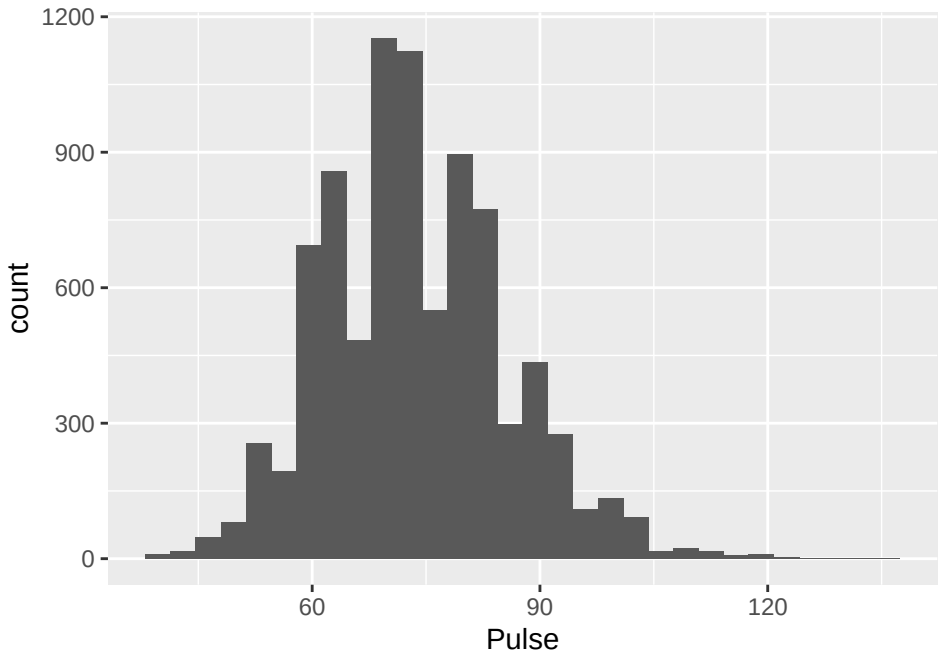

Figure 1: Frequency Distribution of (60-sec) Pulse

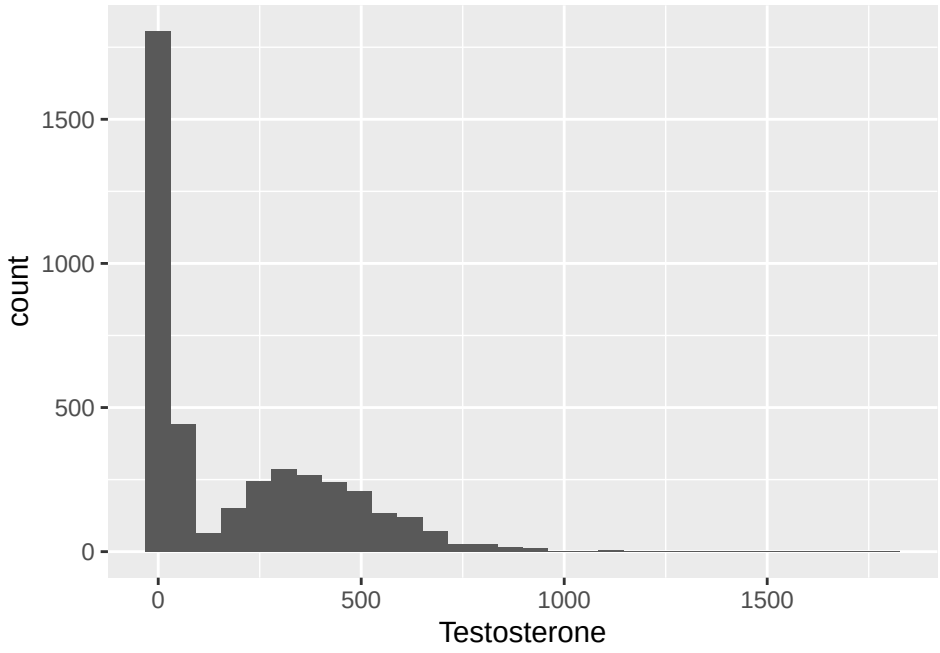

Figure 2: Frequency Distribution for Testosterone Levels (ng/ml)

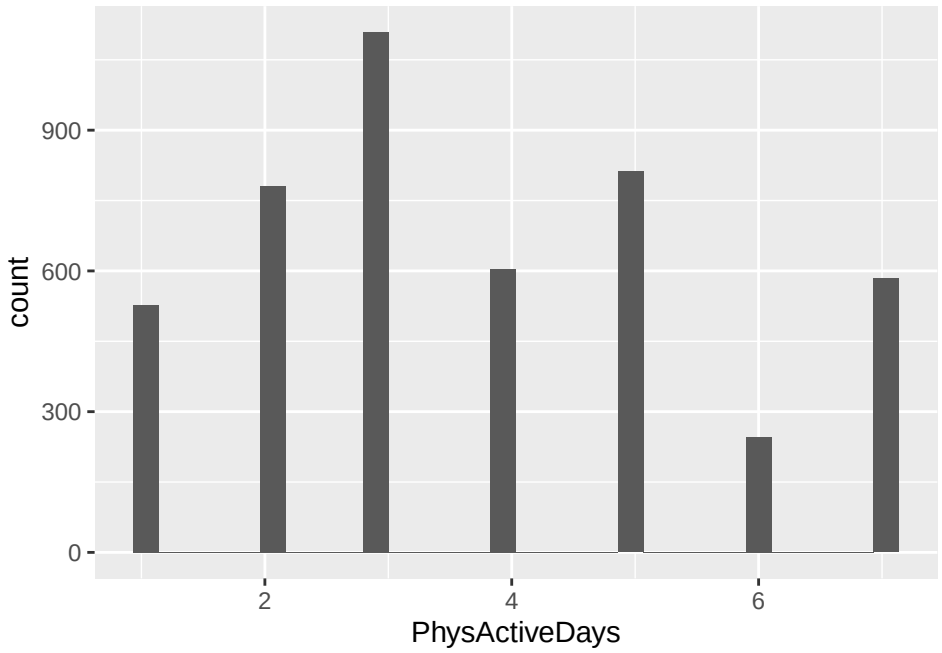

Figure 3: Frequency Distribution for Physically Active Days (per week)

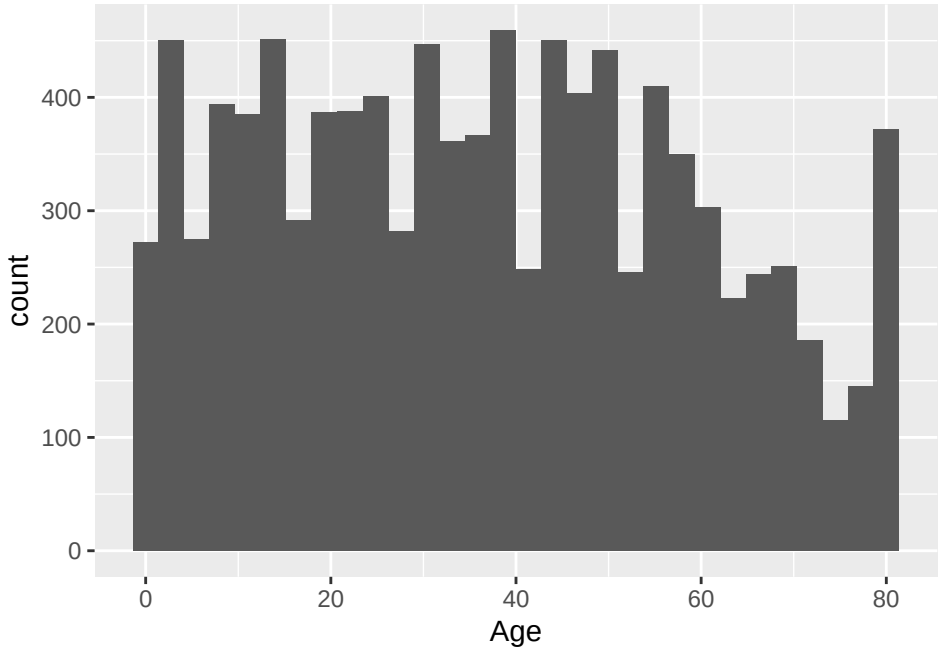

Figure 4: here is the caption

Figure 5: Frequency of Males vs. Females

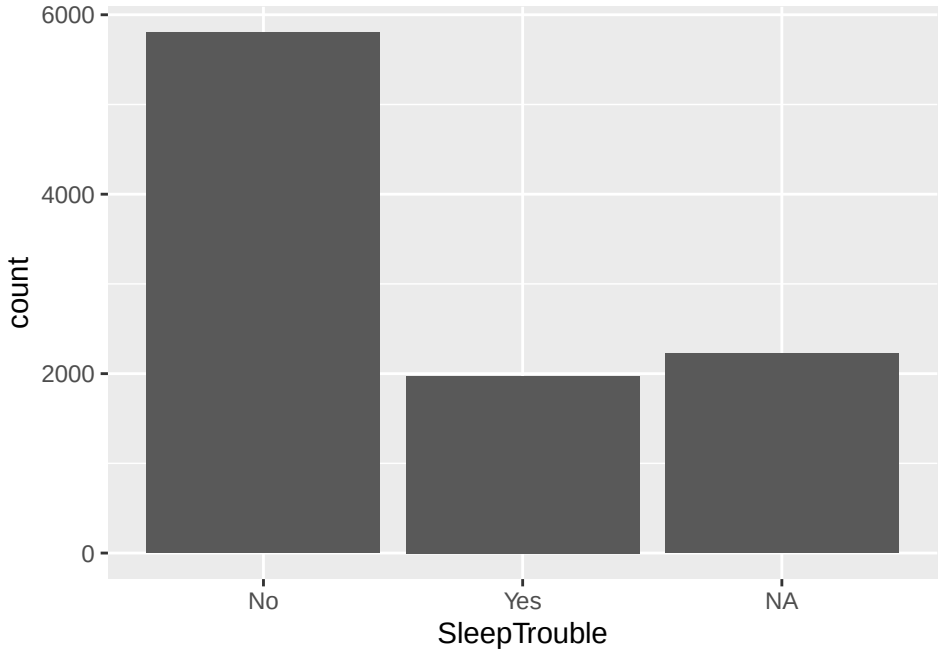

Figure 6: Frequency of if Sleep Trouble reported to Medical Staff

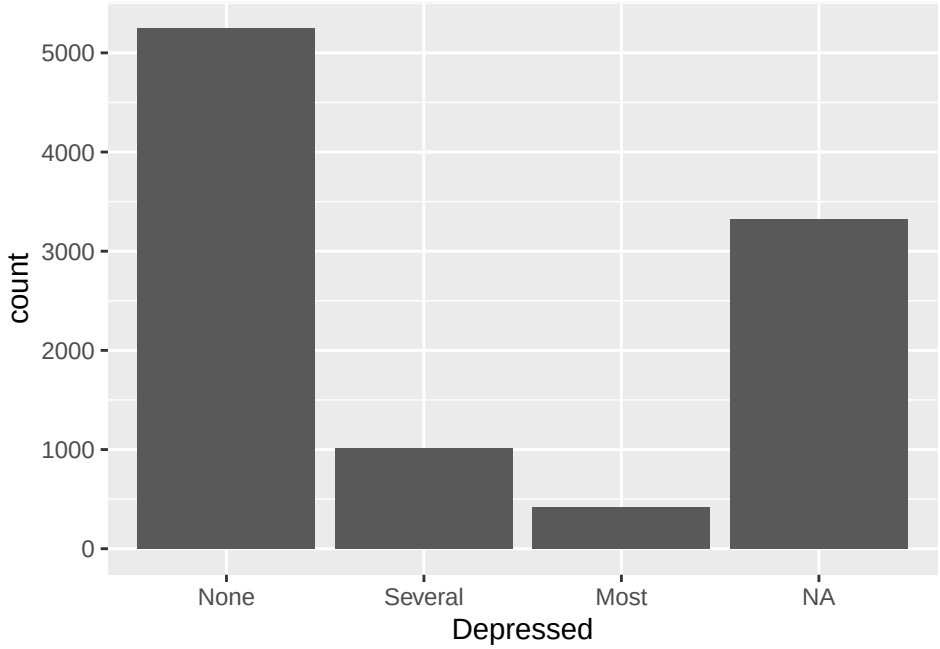

Figure 7: Frequency of Depression (based on proportion of days/week)

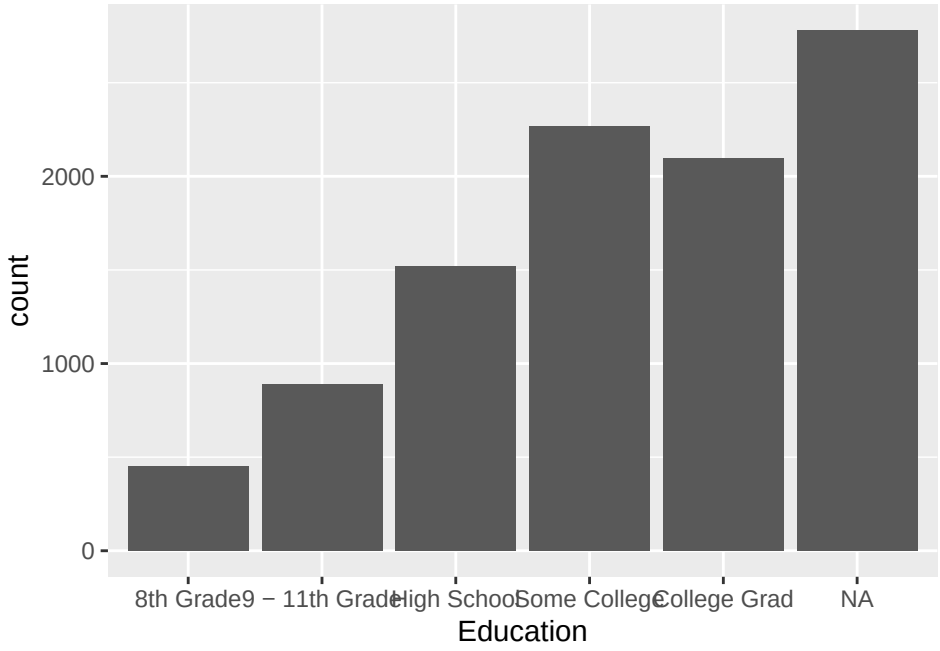

Figure 8: Frequency for Proprotion of Education Completed

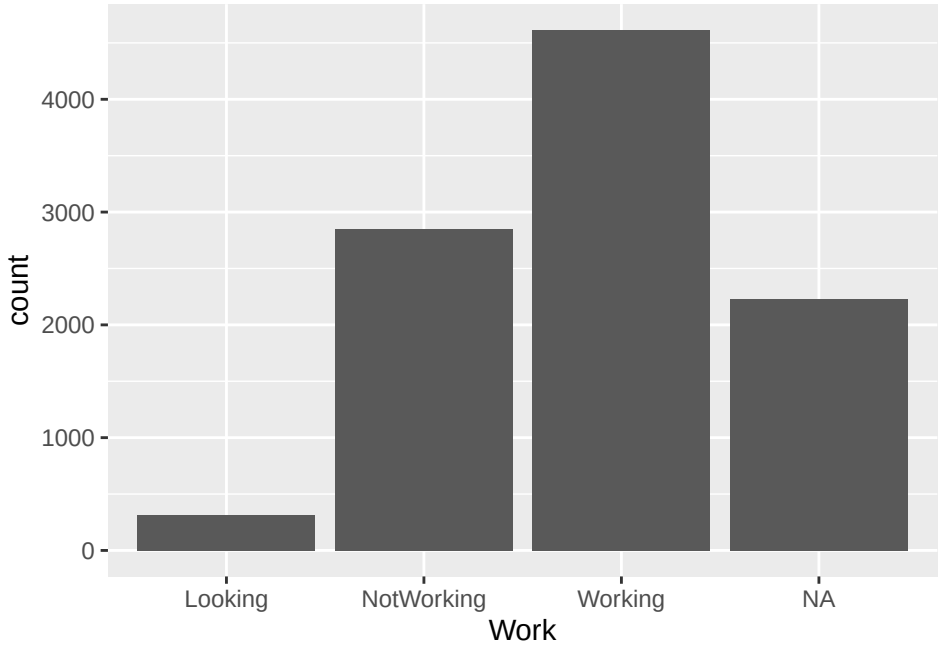

Figure 9: Frequency of Current Work Status

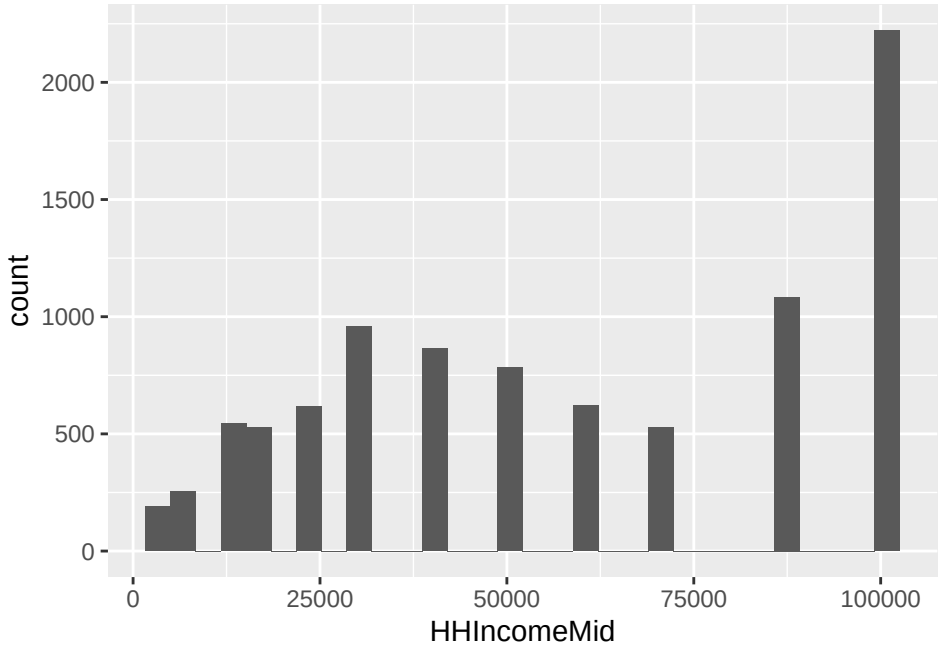

Figure 10: Frequency of Household Income Salaries

```
## Rows: 10,000
## Columns: 10
## $ Age            <int> 34, 34, 34, 4, 49, 9, 8, 45, 45, 45, 66, 58, 54, 10, 58~
## $ HHIncomeMid    <int> 30000, 30000, 30000, 22500, 40000, 87500, 60000, 87500,~
## $ Testosterone   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ PhysActiveDays <int> NA, NA, NA, NA, NA, NA, NA, 5, 5, 5, 7, 5, 1, NA, 2, 7,~
## $ Pulse          <int> 70, 70, 70, NA, 86, 82, 72, 62, 62, 62, 60, 62, 76, 80,~
## $ Gender         <fct> male, male, male, male, female, male, male, female, fem~
## $ SleepTrouble   <fct> Yes, Yes, Yes, NA, Yes, NA, NA, No, No, No, No, No, Yes~
## $ Depressed      <fct> Several, Several, Several, NA, Several, NA, NA, None, N~
## $ Education      <fct> High School, High School, High School, NA, Some College~
## $ Work           <fct> NotWorking, NotWorking, NotWorking, NA, NotWorking, NA,~
```

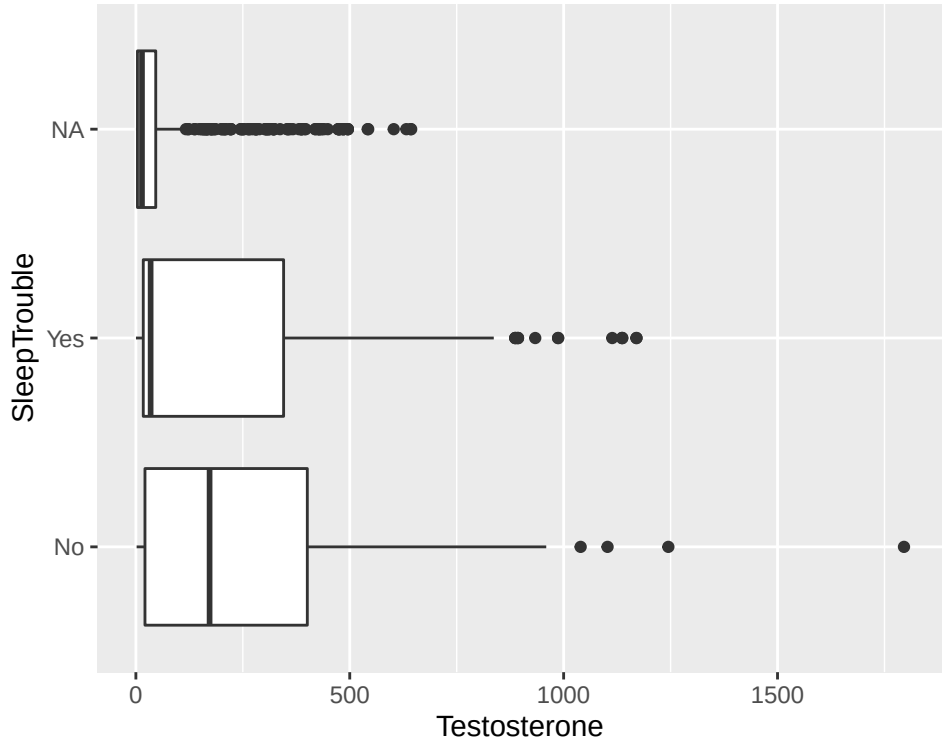

Figure 11: Box Plot of Testosterone Levels (ng/ml) vs. Sleep Troubles


```
## Rows: 10,000
## Columns: 10
## $ Age            <int> 34, 34, 34, 4, 49, 9, 8, 45, 45, 45, 66, 58, 54, 10, 58~
## $ HHIncomeMid    <int> 30000, 30000, 30000, 22500, 40000, 87500, 60000, 87500,~
## $ Testosterone   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ PhysActiveDays <int> NA, NA, NA, NA, NA, NA, NA, 5, 5, 5, 7, 5, 1, NA, 2, 7,~
## $ Pulse          <int> 70, 70, 70, NA, 86, 82, 72, 62, 62, 62, 60, 62, 76, 80,~
## $ Gender         <fct> male, male, male, male, female, male, male, female, fem~
## $ SleepTrouble   <fct> Yes, Yes, Yes, NA, Yes, NA, NA, No, No, No, No, No, Yes~
## $ Depressed      <fct> Several, Several, Several, NA, Several, NA, NA, None, N~
## $ Education      <fct> High School, High School, High School, NA, Some College~
## $ Work           <fct> NotWorking, NotWorking, NotWorking, NA, NotWorking, NA,~
```

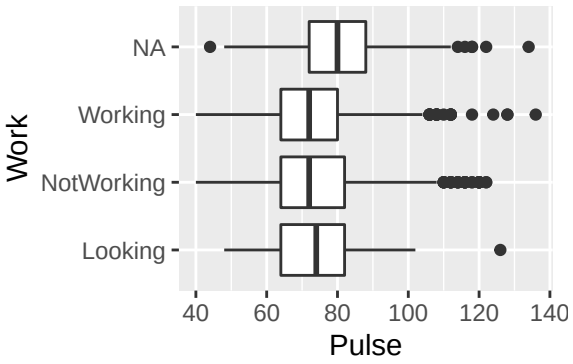

Figure 12: Box Plot of (60 second) Pulse vs. Work

Additional Information

A comment on anything of interest that occurred in doing the project. Were the data approximately what you expected or did some of the results surprise you? How did the sampling go? Do you think you got a representative sample of your population?

Because we will be doing hypothesis testing as the next step, you need to indicate what population your data describes. If it is a census, then maybe it is representative of an even larger population? (For example, a census of state information from 2015 might be somewhat representative of 2016? Is it?) Also, discuss the limitations of describing a larger population

The NHANES data represents a sample from the population of resident citizens of the US (who are 2 months or older). The specific NHANES data we are using was collection for 2009-2010 and 2011-2012. These results were voluntary and primarily collected via interview and physical examinations. There are some limitations of describing the larger population using NHANES data:

Firstly, many of our relevant variables were only collected within adults. Therefore, there are limitations to how we could apply our sample data to the population of all resident citizens of the US (who are 2 months or older) because we do not have certain variable data collected for non-adults.

Additionally, NHANES' study was voluntary and primarily self-reported or collected by healthcare professionals. Therefore, it is likely that people who feel sicker were overrepresented in our sample. Additionally, it is likely that people who have health insurance (or can afford wellness check-ups etc.) are also overrepresented in our sample.