m158-project1-sara-ian

Reponse Variable

Our response variable is going to be pulse which is a quantitative variable.

SLR

pulse versus testosterone

Hypothesis

null: $\beta_1 = 0$

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 these 10 are relevant to our project:

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

Gender (categorical)

- Gender (sex) of study particpant 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.

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).

Filtering NHANES for to contain only our relevant variables:

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

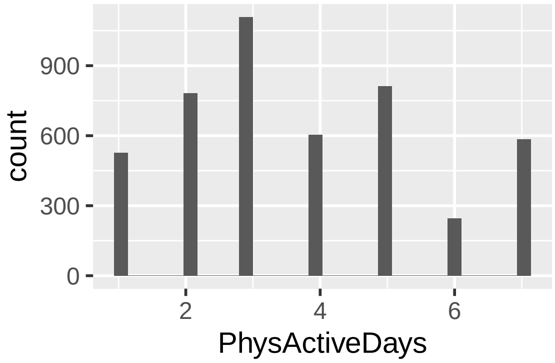
Summary Statistics

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

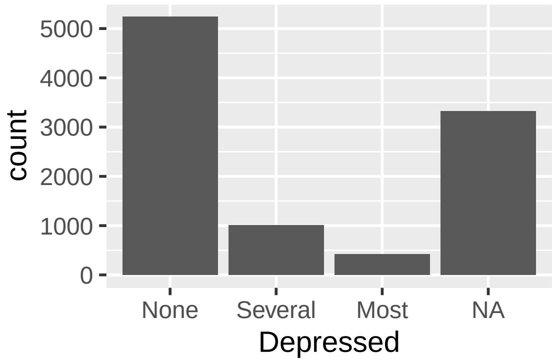
```
## # A tibble: 10 x 5
##    skim_variable  numeric.mean numeric.sd complete_rate numeric.p50
##    <chr>                 <dbl>      <dbl>         <dbl>       <dbl>
##  1 Gender               NA         NA             1           NA
##  2 SleepTrouble         NA         NA             0.777       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
```

Graphs

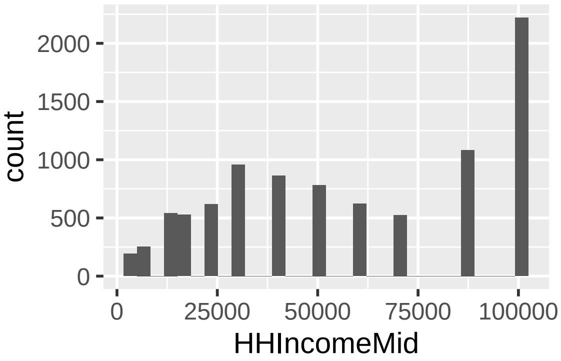
```
OurData %>%
  ggplot(aes(x = PhysActiveDays)) +
  geom_histogram()
```



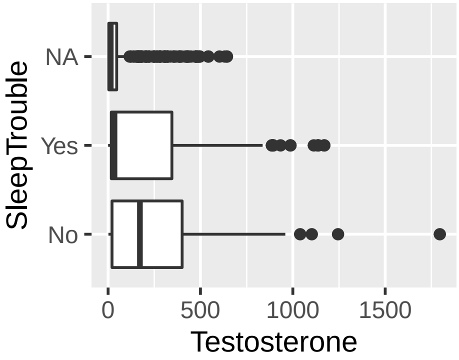
```
OurData %>%
  ggplot(aes(x = Depressed)) +
  geom_bar()
```



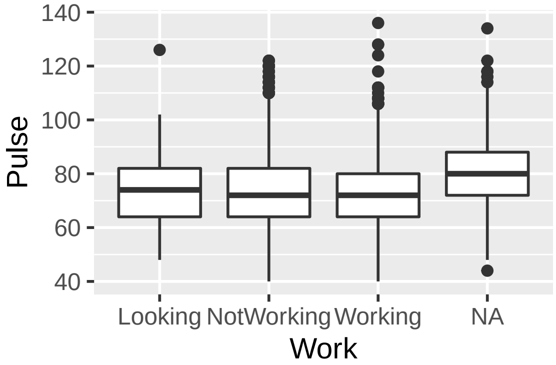
```
OurData %>%
  ggplot(aes(x = HHIncomeMid)) +
  geom_histogram()
```



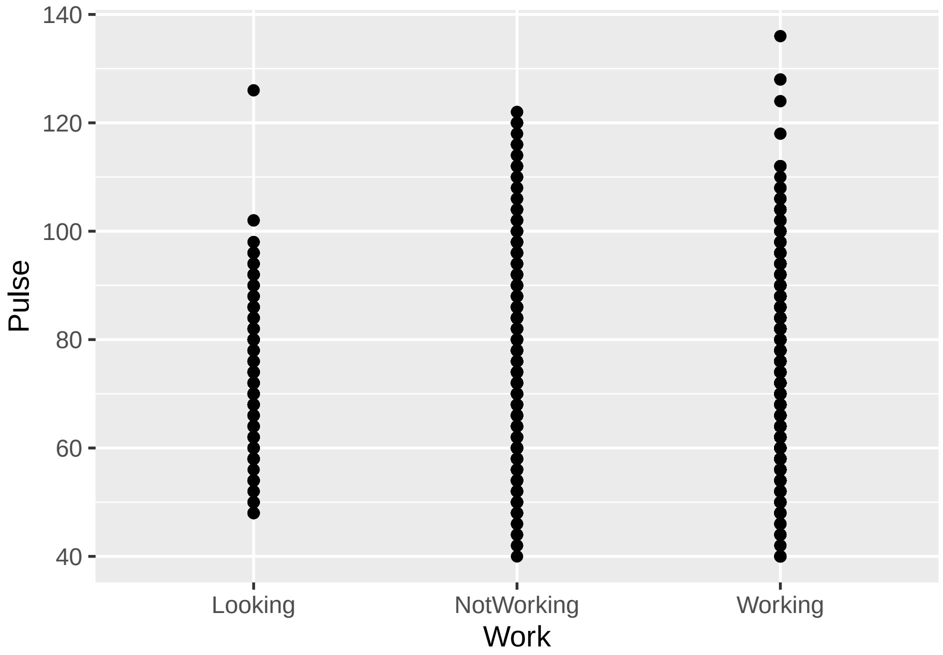
```
ggplot(OurData, aes(x= Testosterone, y = SleepTrouble)) + geom_boxplot()
```



```
ggplot(OurData, aes(x= Work, y = Pulse)) + geom_boxplot()
```



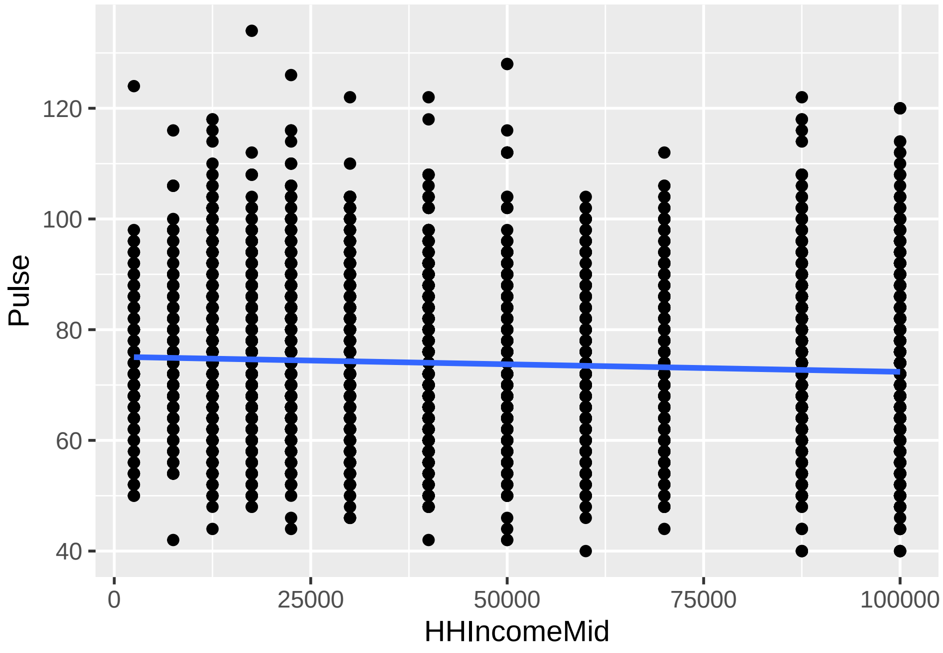
```
OurData %>%
 lm(Pulse ~ Work, data = .) %>%
 ggplot(aes(x=Work, y=Pulse)) +
 geom_point() +
 geom_smooth(method= "lm", se = FALSE)
```



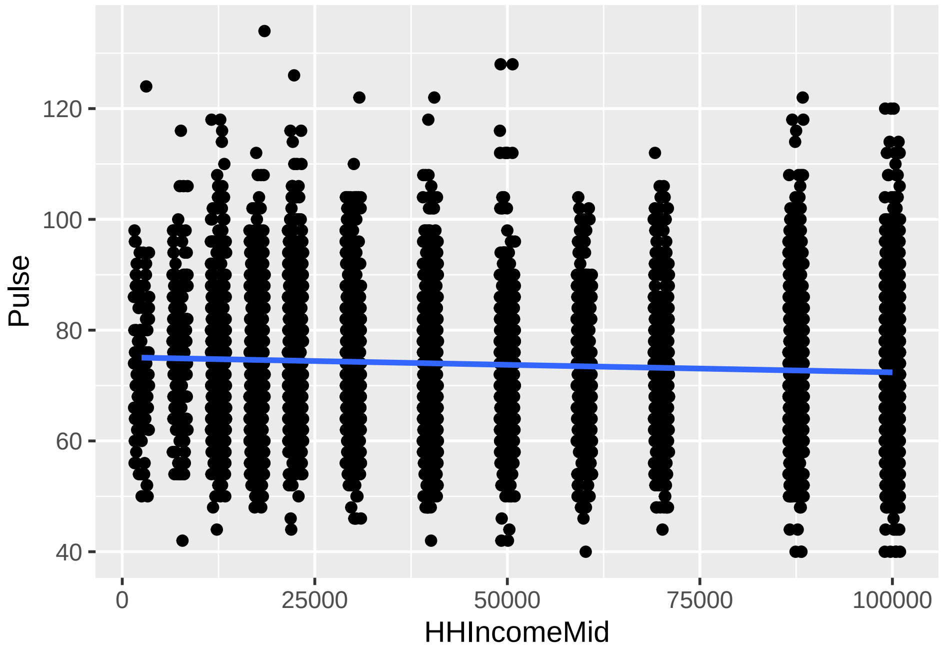
```
OurData %>%
  lm(Pulse ~ Work, data = .) %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term           estimate std.error statistic p.value
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    73.6         0.691  106.      0
## 2 WorkNotWorking  0.00734     0.728    0.0101  0.992
## 3 WorkWorking    -1.55        0.714   -2.17    0.0297
```

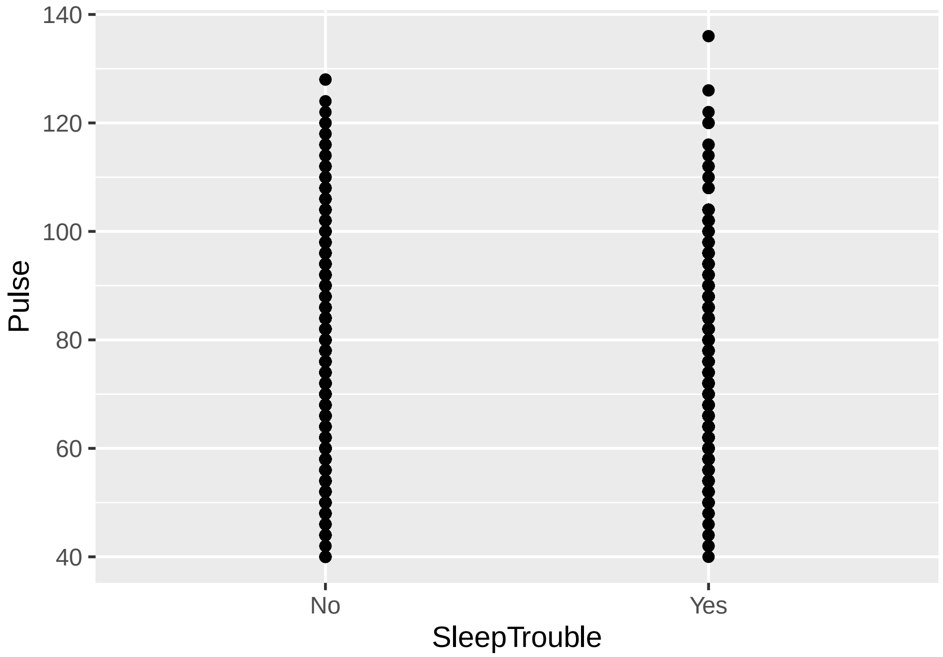
```
OurData %>%
  lm(Pulse ~ HHIncomeMid, data = .) %>%
  ggplot(aes(x=HHIncomeMid, y=Pulse)) +
  geom_point() +
  geom_smooth(method= "lm", se= FALSE)
```



```
OurData %>%
 lm(Pulse ~ HHIncomeMid, data = .) %>%
 ggplot(aes(x=HHIncomeMid, y=Pulse)) +
 geom_jitter(height = 0, width = 1000) +
 geom_smooth(method= "lm", se= FALSE)
```



```
OurData %>%
 lm(Pulse ~ SleepTrouble, data = .) %>%
 ggplot(aes(x=SleepTrouble, y=Pulse)) +
 geom_point() +
 geom_smooth(method= "lm", se= FALSE)
```



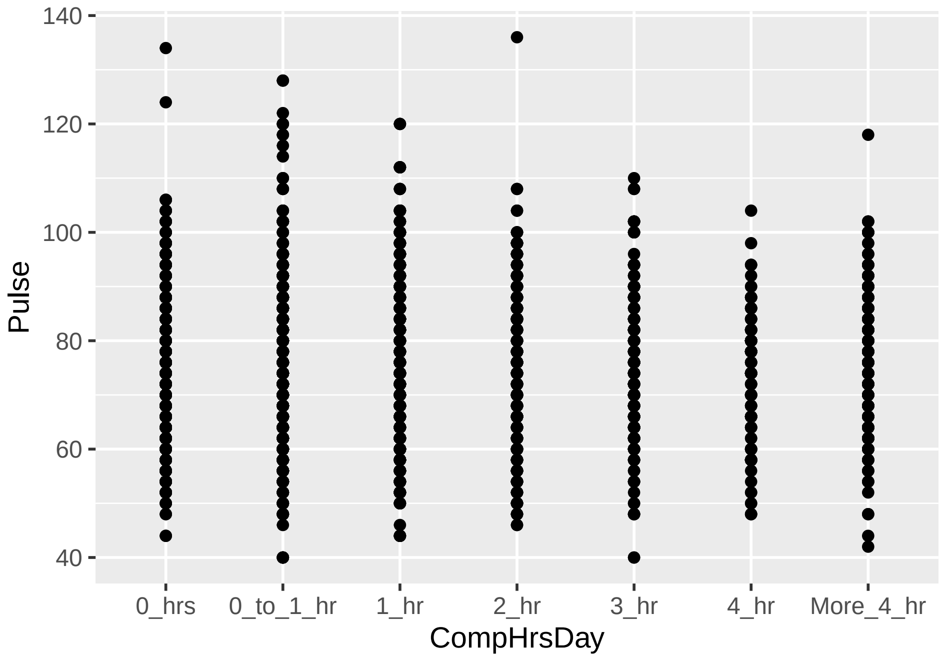
```
OurData %>%
  lm(Pulse ~ SleepTrouble, data = .) %>%
  tidy()
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic p.value
##   <chr>              <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)       72.4       0.160    452.   0
## 2 SleepTroubleYes    0.890     0.317      2.81 0.00497
```

```
 OurData %>%
  lm(Pulse ~ as.factor(HHIncomeMid), data = .) %>%
  tidy()
```

```
## # A tibble: 12 x 5
##    term                          estimate std.error statistic p.value
##    <chr>                            <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)                    73.7         1.02   72.6     0
##  2 as.factor(HHIncomeMid)7500      1.78        1.33    1.34    0.180
##  3 as.factor(HHIncomeMid)12500     1.33        1.16    1.14    0.254
##  4 as.factor(HHIncomeMid)17500    -0.0500      1.17   -0.0428  0.966
##  5 as.factor(HHIncomeMid)22500     2.31        1.15    2.01    0.0441
##  6 as.factor(HHIncomeMid)30000     0.652       1.10    0.592   0.554
##  7 as.factor(HHIncomeMid)40000     0.591       1.11    0.534   0.594
##  8 as.factor(HHIncomeMid)50000    -1.06        1.12   -0.949   0.343
##  9 as.factor(HHIncomeMid)60000    -1.45        1.14   -1.28    0.202
## 10 as.factor(HHIncomeMid)70000     0.106       1.16    0.0919  0.927
## 11 as.factor(HHIncomeMid)87500     0.0613      1.09    0.0563  0.955
## 12 as.factor(HHIncomeMid)100000   -1.56        1.05   -1.49    0.137
```

```
 NHANES %>%
  lm(Pulse ~ CompHrsDay, data = .) %>%
  ggplot(aes(x=CompHrsDay, y=Pulse)) +
  geom_point() +
  geom_smooth(method= "lm", se= FALSE)
```

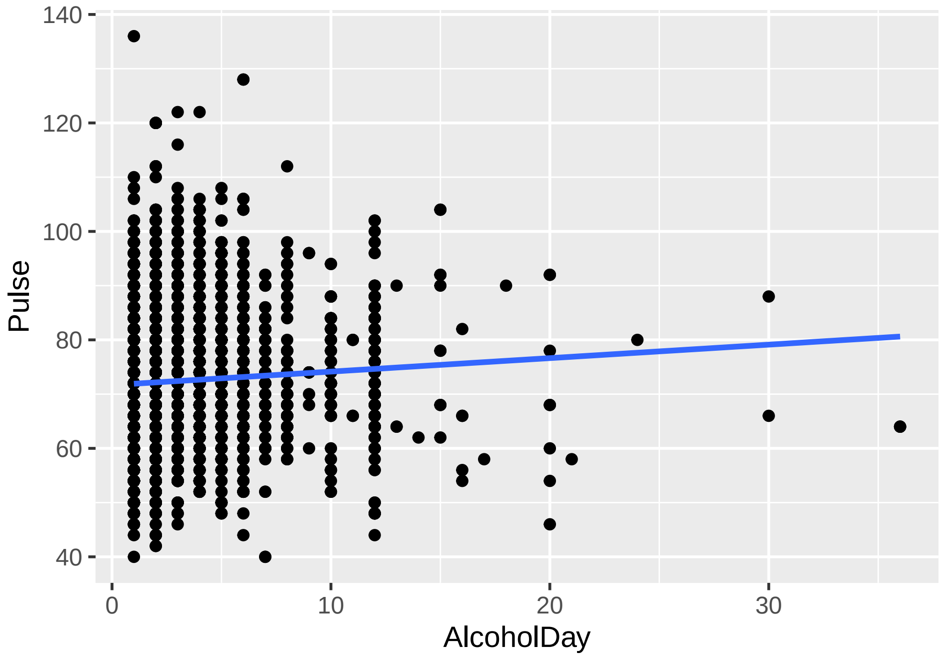


```
 NHANES %>%
  lm(Pulse ~ CompHrsDay, data = .) %>%
   tidy()
```

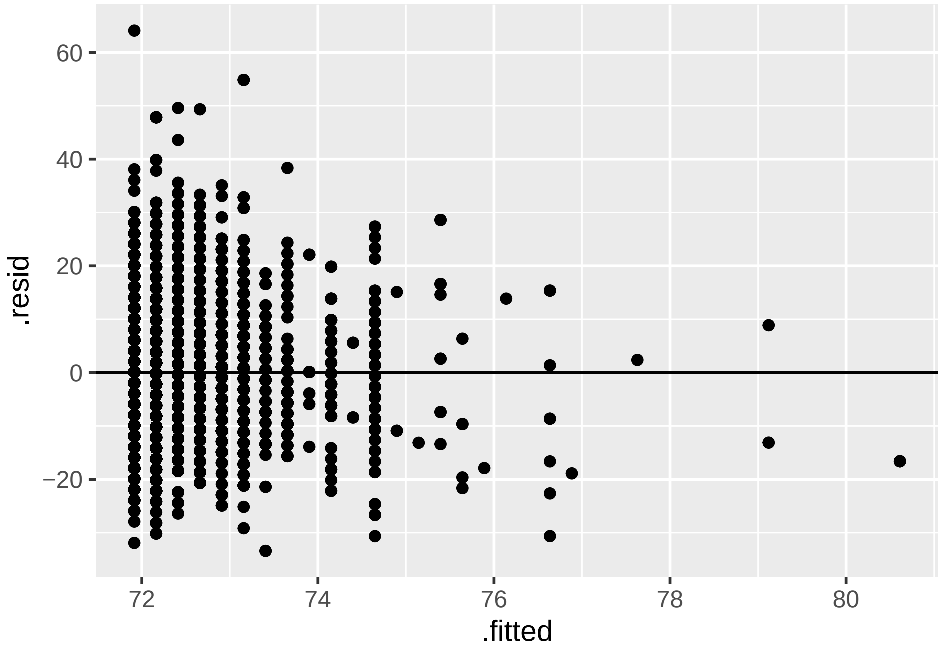
```
## # A tibble: 7 x 5
##   term                 estimate std.error statistic   p.value
##   <chr>                   <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)            72.4       0.415   174.    0
## 2 CompHrsDay0_to_1_hr     1.36      0.540     2.51  0.0121
## 3 CompHrsDay1_hr          1.49      0.575     2.59  0.00952
## 4 CompHrsDay2_hr          0.981     0.663     1.48  0.139
## 5 CompHrsDay3_hr          1.93      0.786     2.45  0.0142
## 6 CompHrsDay4_hr          0.684     1.05      0.648 0.517
## 7 CompHrsDayMore_4_hr     3.79      0.889     4.26  0.0000205
```

```
Alcohol_new <- NHANES %>%
  filter(AlcoholDay <= 40)
```

```
Alcohol_new %>%
  lm(Pulse ~ AlcoholDay, data = .) %>%
  ggplot(aes(x=AlcoholDay, y=Pulse)) +
  geom_point() +
  geom_smooth(method= "lm", se= FALSE)
```



```
Alcohol_new %>%
  lm(Pulse ~ AlcoholDay, data = .) %>%
  augment() %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```

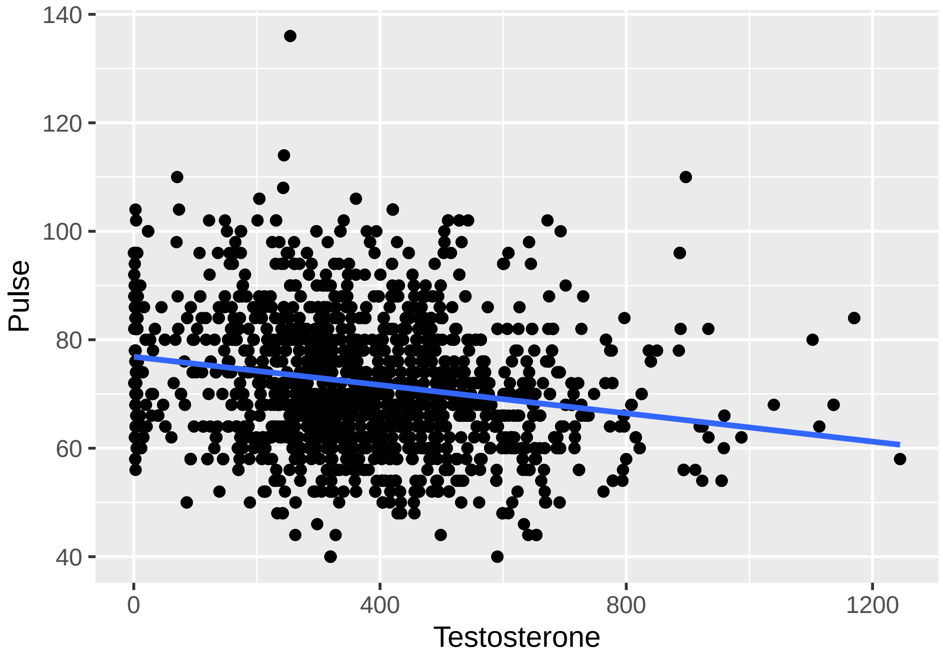


```
 NHANES %>%
  lm(Pulse ~ AlcoholDay, data = .) %>%
   tidy()
```

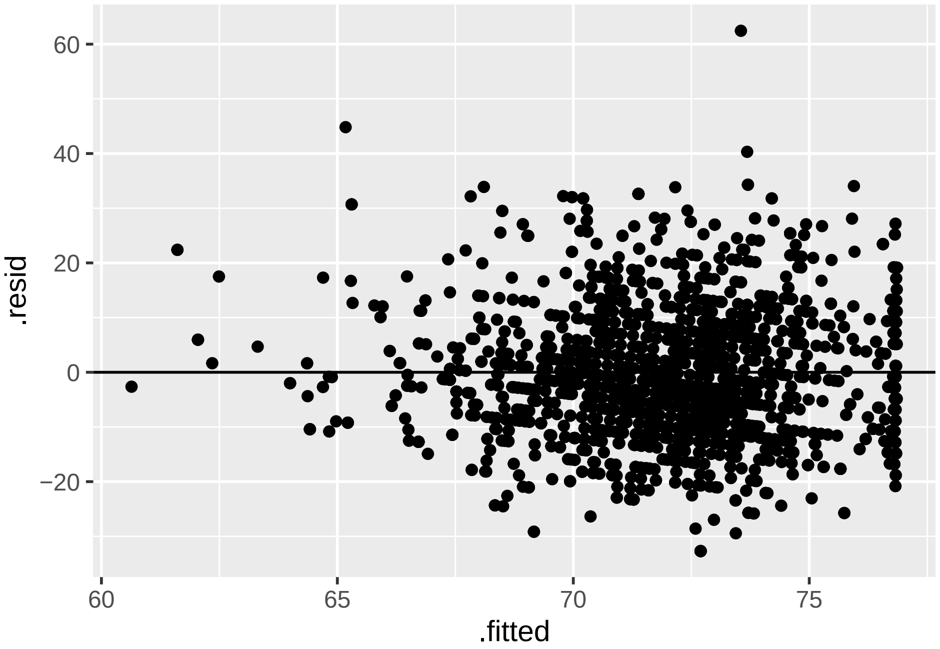
```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   71.8      0.227     317.   0
## 2 AlcoholDay     0.197    0.0526      3.75 0.000177
```

```
testosterone_new <-OurData %>%
  filter(Testosterone < 1400 ) %>%
  filter(Gender == "male")
```

```
 testosterone_new %>%
  lm(Pulse ~ Testosterone, data = .) %>%
  ggplot(aes(x=Testosterone, y=Pulse)) + geom_point() + geom_smooth(method= "lm", se= FALSE)
```



```
 testosterone_new %>%
  lm(Pulse ~ Testosterone, data = .) %>%
  augment() %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



```
testosterone_new %>%
  lm(Pulse ~ Testosterone, data = .) %>%
  tidy(conf.int = TRUE, conf.level = 0.95)
```

```
## # A tibble: 2 x 7
##   term         estimate std.error statistic  p.value conf.low conf.high
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)   76.9      0.586      131.   0          75.7      78.0
## 2 Testosterone  -0.0130   0.00137     -9.52 5.08e-21  -0.0157   -0.0103
```

```
testosterone_new %>%
  lm(Pulse ~ Testosterone, data = .) %>%
  glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl>  <dbl>  <dbl>
## 1    0.0438        0.0433  11.9      90.5 5.08e-21     1 -7694. 15394. 15411.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

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.