Assignment 2

ML

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
gss <- read_csv("https://raw.githubusercontent.com/mjclawrence/soci385_f23/main/data/assig
```

```
Rows: 5622 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): class, region, racehisp
dbl (2): id, hrs1

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

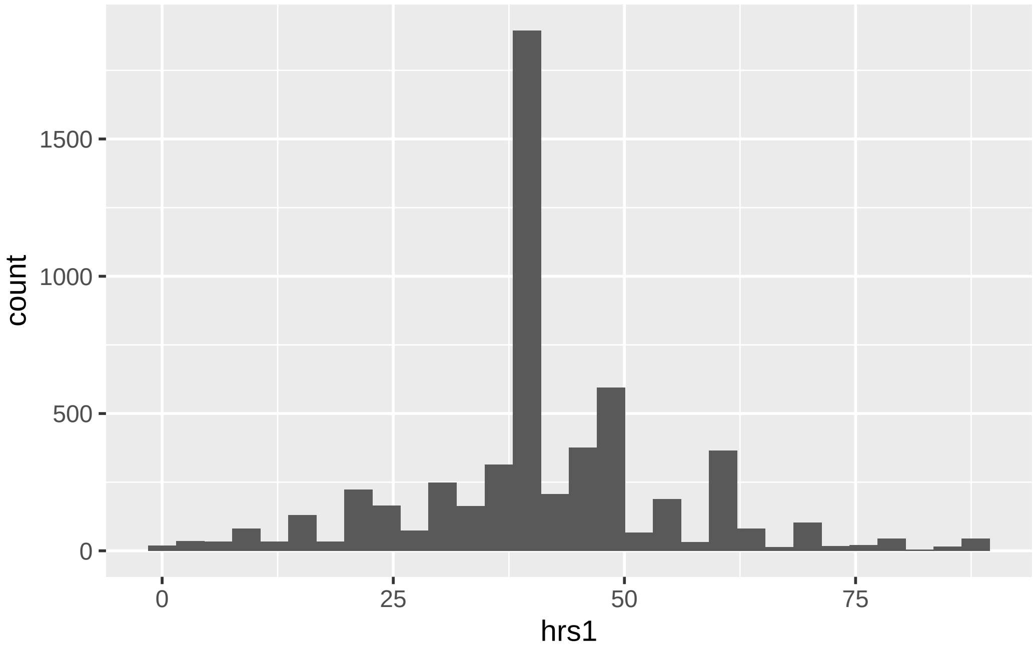
Questions

1. Create a histogram showing the distribution of hours worked. Remember to label your axes and provide a title and subtitle. Use your histogram to briefly summarize the center

and shape of this distribution. Note: don't waste time adjusting the binwidth; you can use the default of 30 here.

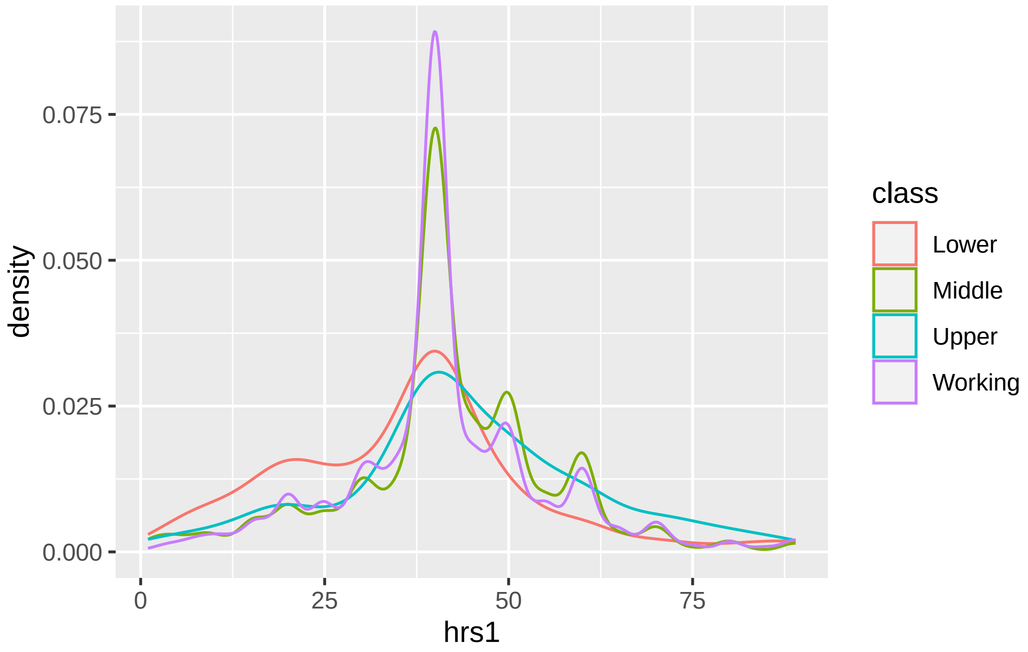
```
gss |>
  ggplot(aes(x = hrs1)) + geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2. Create a figure with overlapping density plots of hours worked for each social class. Use this plot to roughly compare the probabilities that respondents from each class will work 40 hours a week (it's the mode for each class).

```
gss |>
  ggplot(aes(x = hrs1, color = class)) + geom_density()
```



3. Create a new factor variable for hours worked with the following categories: less than 20, 20-39, 40, more than 40. The levels should be in order from least to most hours.

```
gss <- gss |>
  mutate(hours = ifelse(hrs1 < 20, "Less than 20",
                        ifelse(hrs1 %in% 20:39, "20-39",
                               ifelse(hrs1 == 40, "40",
                                      "More than 40"))),
         hours = factor(hours,
                        levels = c("Less than 20",
                                   "20-39",
                                   "40",
                                   "More than 40")))
```

4. Using the new variable from #3, create a table showing the proportion of respondents in each region who are in each category of hours worked. Which region has the highest proportion of respondents who work more than forty hours a week?

```
round(prop.table(table(gss$region, gss$hours),1),2)
```

```
               Less than 20 20-39   40 More than 40
```

```
  Middle Atlantic         0.08  0.23 0.32         0.37
  Midwest                 0.06  0.21 0.31         0.42
  New England             0.07  0.26 0.30         0.36
  South                   0.06  0.25 0.33         0.36
  West                    0.08  0.23 0.32         0.37
```

```r
library(kableExtra)
```

```

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```r
round(prop.table(table(gss$region, gss$hours),1),2) |>
  kable(booktabs = TRUE, align = rep('c', 4))
```

	Less than 20	20-39	40	More than 40
Middle Atlantic	0.08	0.23	0.32	0.37
Midwest	0.06	0.21	0.31	0.42
New England	0.07	0.26	0.30	0.36
South	0.06	0.25	0.33	0.36
West	0.08	0.23	0.32	0.37

5. Find the standard deviation of hours worked by race/ethnicity in the New England region. In a sentence, describe any similarities or differences you notice. You do not have to create a figure for this question.

```r
gss |>
  filter(region == "New England") |>
  group_by(racehisp) |>
  summarise(sd_hours = sd(hrs1))
```

```
# A tibble: 4 x 2
  racehisp sd_hours
  <chr>       <dbl>
1 Black        9.53
2 Hispanic    14.7
```

```
3 Other        16.5
4 White        13.6
```

6. Approximately how long did it take you to complete this assignment?