Problem Set 1

Due date: 18 September

Please upload your completed assignment to the ELMs course site (under the assignments menu). Remember to include an annotated script file for all work with R and show your math for all other problems (if applicable, or necessary). Please also upload your completed assignment to the Github repository that you have shared with us. *We should be able to run your script with no errors.*

Total points: 25

Ans {

Initializing the R script

```
#Disable warnings globally
options(warn=-1)
#.rs.restartR()

#Package loading
requiredPackages <- c("poliscidata",
                      "ggplot2",
                      "magrittr",
                      "dplyr",
                      "data.table")

#Installs packages if not yet installed
for (package in requiredPackages)
{
  if (!requireNamespace(package, quietly = TRUE))
  install.packages(package)

}
```

```
Registered S3 method overwritten by 'gdata':
  method         from
  reorder.factor gplots
```

```
  invisible(lapply(requiredPackages, require, character.only = TRUE, quietly = T))
```

```
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


Attaching package: 'data.table'

The following objects are masked from 'package:dplyr':

    between, first, last
```

```
  rm(list=ls())
}
```

Question 1

Points: 5

Using the `gss` dataset (which has survey data with the individual respondent as the unit of analysis), create a frequency distribution and bar chart for each of the following variables: `degree` and `partyid_3`. Describe the distribution of each variable in detail.

> **Note**
>
> The `gss` dataset can be found in `poliscidata::gss`. Take a look at `?gss` to see what these variables measure.

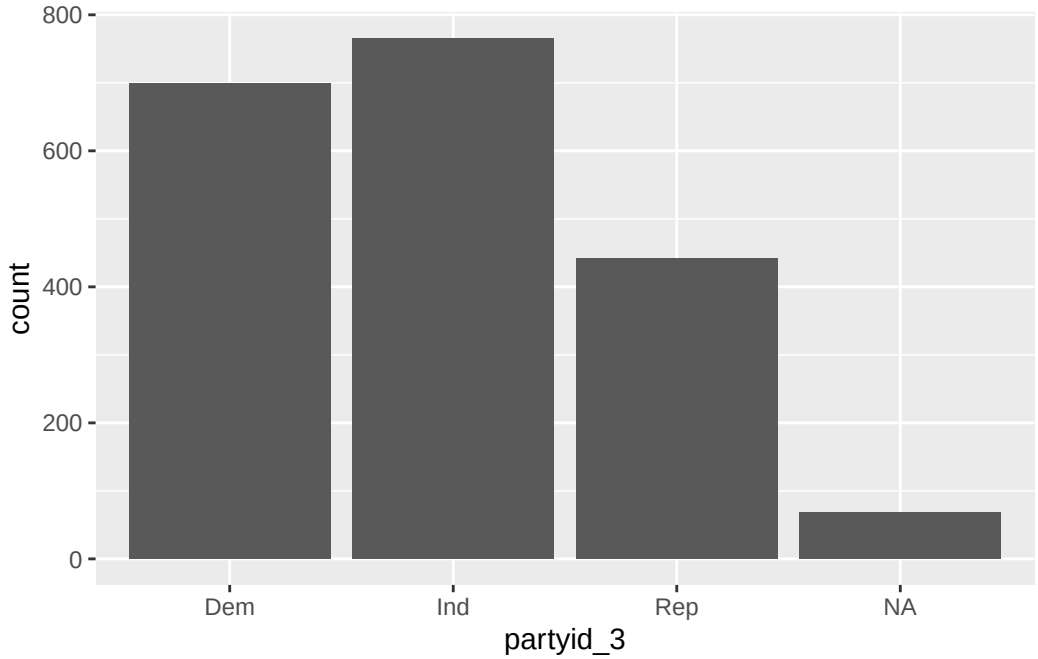
Ans

```
gss_data <-poliscidata::gss

#Exploring the variables
?gss
```

```
#Frequency Dist. for partyid_3
pid_3 <- as.data.frame(summary(gss_data$partyid_3)) %>% setNames(c("Frequencies"))

pid_3 <- pid_3 %>%
                  mutate(
                    Freq_Perc = round(prop.table(pid_3)[,1] * 100, 2),
                          cum = cumsum(Freq_Perc)
                        )
pid_3
```

```
     Frequencies Freq_Perc   cum
Dem          699     35.41 35.41
Ind          765     38.75 74.16
Rep          442     22.39 96.55
NA's          68      3.44 99.99
```

```
#Barplot for partyid_3
g_partyid_3 <- ggplot(gss_data, aes(partyid_3))
g_partyid_3 + geom_bar()
```

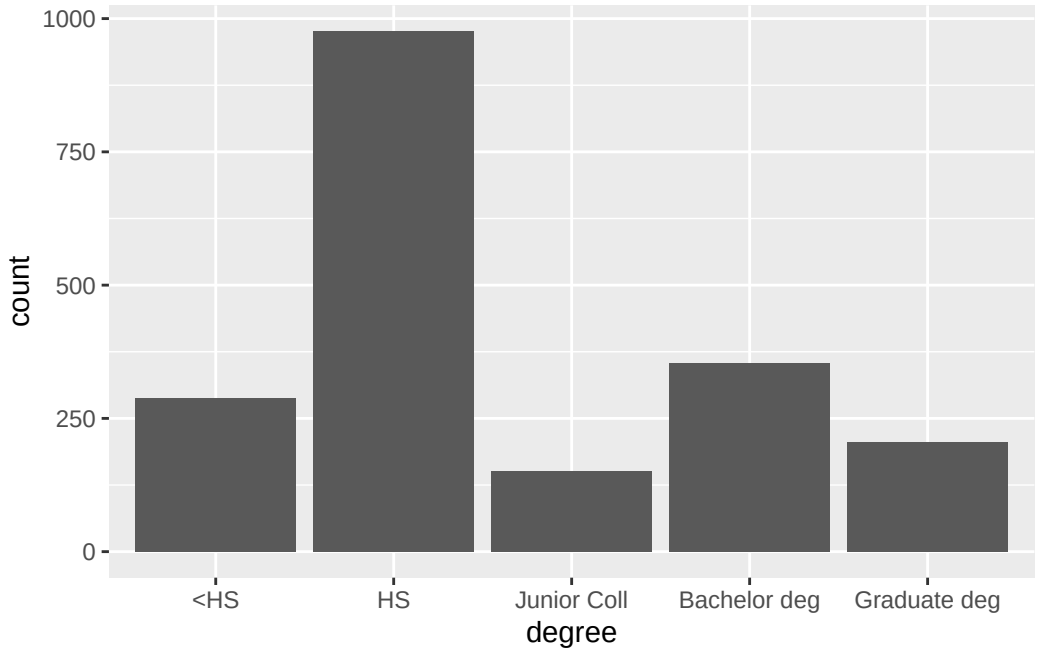


The variable partyid_3 has 1906 values out of which 3 are distinct and 68 are missing. It has three categories, Dem, Ind and Rep with NAs. Ind takes about 39% of all observations followed by Dem at 35% and Rep at 22%. Na's are about 3.4% of the data.

```
#Frequency Dist. for degree
d2 <- as.data.frame(summary(gss_data$degree)) %>% setNames(c("Frequencies"))

d2 <- d2 %>%
            mutate(
              Freq_Perc = round(prop.table(d2)[,1] * 100, 2),
                    cum = cumsum(Freq_Perc)
                  )
d2
```

```
             Frequencies Freq_Perc    cum
<HS                  288     14.59  14.59
HS                   976     49.44  64.03
Junior Coll          151      7.65  71.68
Bachelor deg         354     17.93  89.61
Graduate deg         205     10.39 100.00
```

```r
#Bar chart for degree
g_degree <- ggplot(gss_data, aes(degree))
g_degree + geom_bar()
```



The variable degree has 1974 values out of which 5 are distinct and 0 are missing. It has five categories, <HS, HS and Junior Coll, Bachelor deg and Graduate deg. HS takes about 49% of all observations followed by Bachelor at 18% and <HS at 14%. Graduate is at 10% and Junior Coll is at 8%.

Question 2

Examine the following vector of (fake) student IQ test scores:

```r
iq <- c(
  145, 139, 126, 122, 125, 130, 96, 110, 118, 118, 101, 142, 134, 124, 112, 109,
  134, 113, 81, 113, 123, 94, 100, 136, 109, 131, 117, 110, 127, 124, 106, 124,
  115, 133, 116, 102, 127, 117, 109, 137, 117, 90, 103, 114, 139, 101, 122, 105,
  97, 89, 102, 108, 110, 128, 114, 112, 114, 102, 82, 101
)
```

Part A

Points: 5

Find the five-number summary, mean, and standard deviation for these data. Also, are there any suspected outliers in the distribution? If so, what are they and how do you know?

The five number summary:

```
#Five number summary
summary(iq)[-4]
```

```
   Min. 1st Qu.  Median 3rd Qu.    Max.
  81.00  104.50  114.00  125.25  145.00
```

The mean:

```
summary(iq)[4]
```

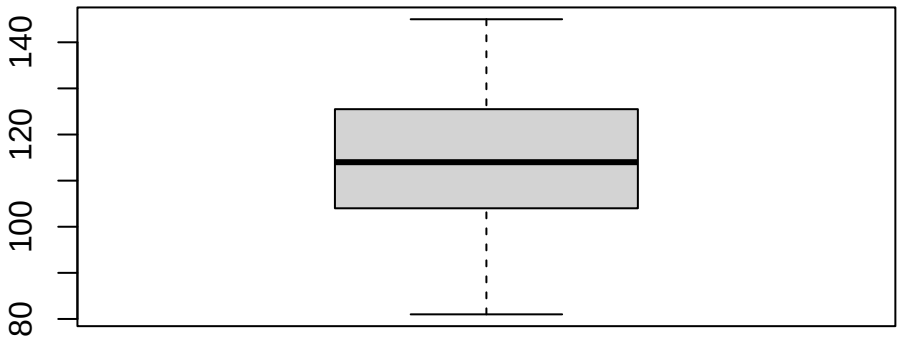
```
    Mean
114.9833
```

The standard deviation:

```
sd(iq)
```

```
[1] 14.80093
```

```
#Outliers:
boxplot(iq)
```



Based upon the boxplot there are no easily identifiable outliers from this dataset.

Part B

Points: 5

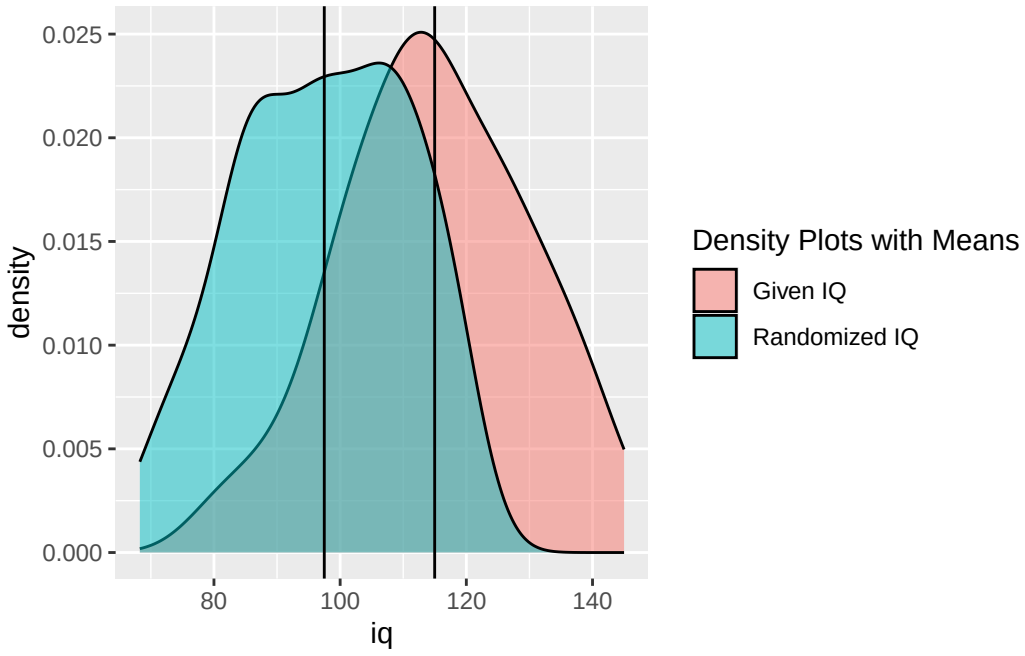
In large populations, IQ scores are standardized to have a mean of 100 and a standard deviation of 15. In what way does the distribution among these students differ from the overall population?

```
mean <- 100
Sd <- 15
```

```
rnorm_fixed <- function(n, mean, sd) {    # Create user-defined function
  as.vector(mean + sd * scale(rnorm(n)))
}

library(truncnorm) #Random normal distb.
x2 = truncnorm::rtruncnorm(100, a=-80, b=120, mean=100, sd=15)

iq_1 <- cbind(as.data.frame(iq), 1)
```

```r
x2_1 <- cbind(as.data.frame(x2), 2)

names(iq_1) <- c('iq', 'group')
names(x2_1) <- c('iq', 'group')

density_data <- rbind(iq_1, x2_1)


density_data$group <- as.character(density_data$group)
```

```r
p<- ggplot(density_data, aes(x = iq, fill = group)) +
      geom_density(alpha = 0.5) +
      geom_vline(xintercept=mean(iq_1$iq)) +
      geom_vline(xintercept=mean(x2_1$iq)) +
      scale_fill_discrete(name="Density Plots with Means",
      labels=c("Given IQ", "Randomized IQ"))
p
```



As we can see from the plots above; if we look at the density plot for a random sample of IQ, the mean is clearly less than the average iq of the participants. This does not mean that the IQ of all the participants will be higher than the average, just that on average the participants

will have a higher IQ. As the standard deviations are close to each other , we can see that the spread of data are also close to one another. This means that the number of values higher and lower than the mean in both distributions are close to the same.

Question 3

Points: 5

A polling firm is interested in determining how different characteristics of an individual affect vote choice. They record the following characteristics for each survey respondent - race, age, household income, and party affiliation. What "type" of variable does each characteristic most likely represent?

Ans:

Race: Nominal

Age: Ordinal

Household Income: Continuous

Party Affiliation: Nominal

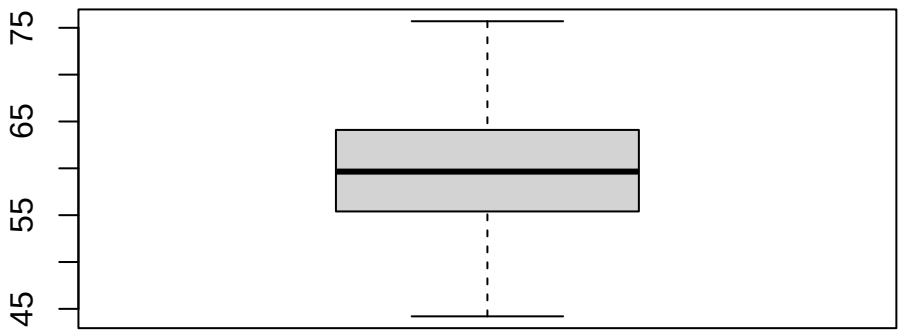
Question 4

Points: 5

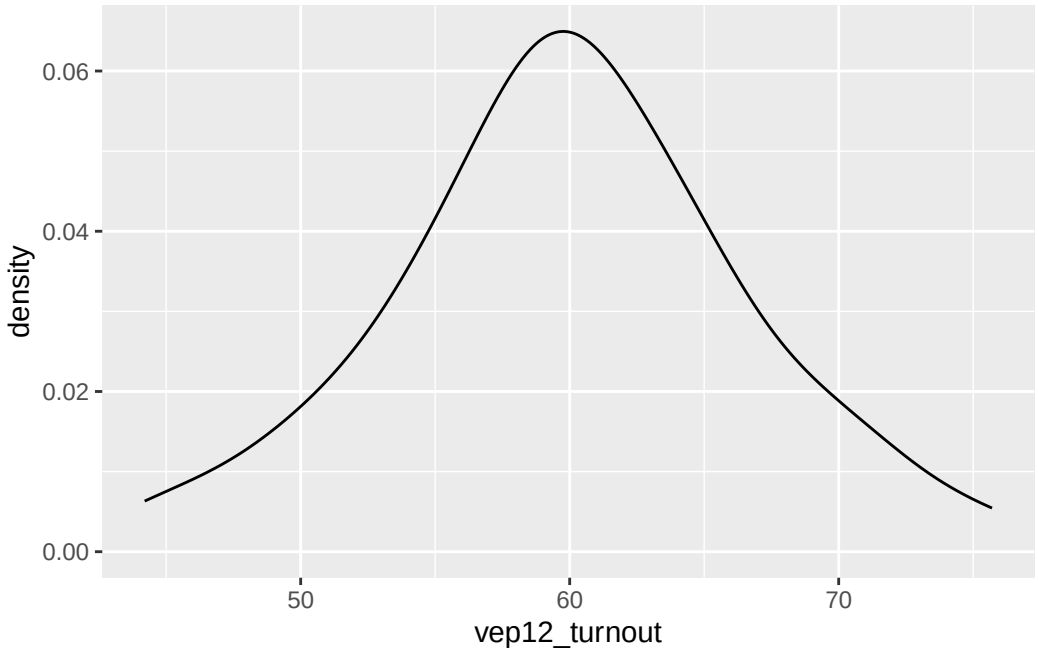
Using the `states` dataset (the unit of analysis is each U.S. state), generate a box plot and density curve for each of the `vep12_turnout` and `cig_tax12` variables. Describe the distribution of each variable in detail.

> **Note**
>
> The `states` data set can be found in `poliscidata::states`. Take a look at `?states` to see what these variables measure.

```
boxplot(poliscidata::states$vep12_turnout)
```

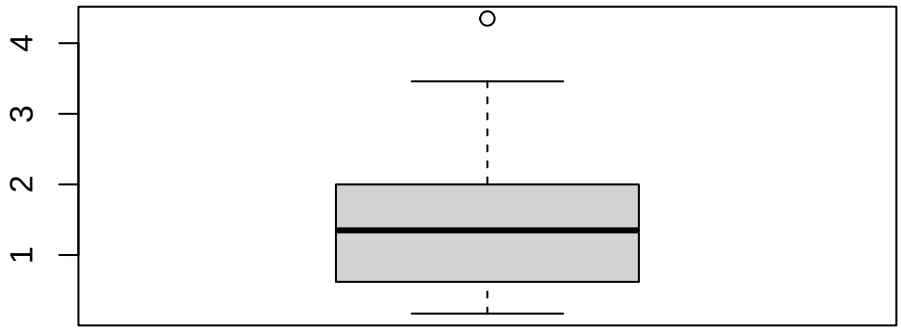



```
ggplot(data = poliscidata::states, aes(x = vep12_turnout)) +
  geom_density()
```

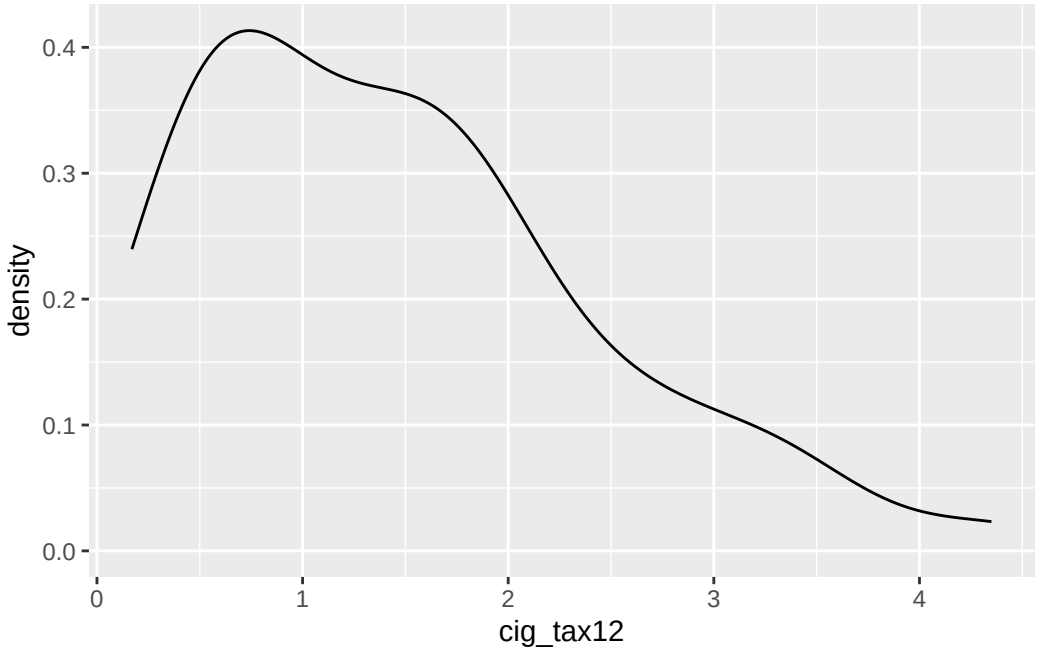


The variable vep12_turnout has 50 values out of which 44 are distinct. It has a minimum value of 44.2, maximum of 75.70 and a median of 59.65. The distribution is single peaked with an easy to define center and equally spaced tails. It has a range of 8.25 and hence is not highly variable. It has no easily identifiable outliers based on IQR.

```
boxplot(poliscidata::states$cig_tax12)
```




```
ggplot(data = poliscidata::states, aes(x = cig_tax12)) +
  geom_density()
```



The variable cig_tax12 has 50 values out of which 41 are distinct. It has a minimum value of 0.172, maximum of 4.350 and a median of 1.349. The distribution is right skewed with two peaks. It has a range of 1.375 and hence is not highly variable. It has a single identifiable outlier of 4.35 that pulls the distribution.