Mid-Term Exam #1

Due date: 6 October (11:59 pm)

Please read all of the questions and tables carefully and follow all instructions. Each question has an allotted point value. Be as thorough and specific as possible; extra calculations and incorrect information, even in the presence of correct information, will result in point deductions. Be sure to show all formulas and all necessary work to answer the questions. You may upload your completed exam to the Elms course site (attach to Exam #1).

> **Note**
>
> While this is an open-note exam, you are not to receive assistance from anyone else. As usual, the Honor Code applies.

```
library(data.table)
library(modelsummary)
library(ggplot2)
```

Conceptual Questions

Total points: 15

Please include all work (and computations) necessary to answer the questions.

Question 1

1 point

The following is a list of observed values, ordered from lowest to highest: 62, 63, 63, 64, 66, 67, 68, 68, 69, 70, 71, 72, 72, 74, 75, 76. What is an accurate five-number summary for these data?

Ans 1:

```r
#The five number summary
Number_values <- c(62, 63, 63, 64, 66, 67, 68, 68, 69, 70, 71, 72, 72, 74, 75, 76)
summary(Number_values)[-4]
```

```
Min. 1st Qu.  Median 3rd Qu.    Max.
62.0    65.5    68.5    72.0    76.0
```

Question 2

1 point

Suppose that the median, Q1, and Q3 from question #1 is an accurate representation of a second (hypothetical) distribution. Based on what this information tells you about this second distribution, which of the following numbers would be a suspected outlier?

A. 76
B. 62
C. 83
D. Both (A) and (C)
E. All of the above
F. None of the above, or cannot be determined from the information given.

Ans 2:

Based on box plot outlier treatment, we can take any value 1.5 times the IQR above the third quartile or 1.5 times the IQR below the second quartile as outliers.

Above outlier:

```r
#IQR = Q3 - Q1

Q1 = 65.5
Q3 = 72.0

IQR = Q3 - Q1

#Hence the IQR is:
IQR
```

```
[1] 6.5
```

```
Lower_bound = Q1 - 1.5 * IQR
Upper_bound = Q3 + 1.5 * IQR

if((76 < Lower_bound) || (76 > Upper_bound))
{
  print('76 is an outlier')
}

if((62 < Lower_bound) || (62 > Upper_bound))
{
  print('62 is an outlier')
}

if((83 < Lower_bound) || (83 > Upper_bound))
{
  print('83 is an outlier')
}
```

```
[1] "83 is an outlier"
```

Based on the above information the correct option is:

C. 83 is higher than the upper bound and hence is a suspected outlier

Question 3

1 point

There is a group of three children with the following ages: 3, 4, and 5. If a 6-year-old child joins the group, what will happen to the mean and standard deviation of the group's age?

```
Previous_mean <- mean(c(3, 4, 5))
Previous_sd <- sd(c(3, 4, 5))

New_mean <- mean(c(3, 4, 5, 6))
New_sd <- sd(c(3, 4, 5, 6))

Mean_sd_df <- data.frame(matrix(ncol = 2, nrow = 2))
colnames(Mean_sd_df) <- c('Before_six_year_old_addition', 'After_six_year_old_addition')
rownames(Mean_sd_df) <- c('Mean', 'Standard_Deviation')

Mean_sd_df[1,1] <- Previous_mean
```

```
  Mean_sd_df[2,1] <- Previous_sd
  Mean_sd_df[1,2] <- New_mean
  Mean_sd_df[2,2] <- New_sd

  Mean_sd_df
```

```
                    Before_six_year_old_addition After_six_year_old_addition
Mean                                           4                    4.500000
Standard_Deviation                             1                    1.290994
```

Hence with the inclusion of the new 6 year old child , both the mean and the standard deviation increases.

Question 4

1 point

If I estimate an OLS regression and obtain a R^2 of 0.40 where the Total Sum of Squares of 4,150, what does the Residual Sum of Squares equal?

Ans 4:

The R2 value can also be calculated as 1 - RSS/TSS

```
  #Hence RSS can be written as:
  RSS = (1 - 0.40) * 4150

  print( paste("RSS value is: ", RSS))
```

```
[1] "RSS value is:  2490"
```

Question 5

1 point

The distribution of some variable has a median that is smaller than its mean. Which of the following statements about the distribution's shape is most consistent with this information?

A. The shape of the distribution would be symmetric
B. The shape of the distribution would be skewed left
C. The shape of the distribution would be skewed right

D. None of the above – cannot be determined from the information given.

Ans 5:

B. The distribution is more skewed to the left

Question 6

1 point

Suppose I want to test the hypothesis that the U.S. public's approval of the president is higher when people have more positive perceptions of the U.S. economy. To test this hypothesis, I conduct a survey of individual-level attitudes. Which of the following research design strategies should I expect to exhibit both the greatest sampling variability (in the context of repeated sampling) and the lowest degree of expected sampling bias?

A. A random sample of 1,200 people from a list of all U.S. residential addresses
B. A random sample of 600 people from a list of all registered students at the University of Maryland
C. A random sample of 1,200 people from a list of all registered students at the University of Maryland
D. A random sample of 600 people from a list of all U.S. residential addresses.

Ans 6:

For this sort of hypothesis testing we require a sample size that is drawn from the entire US population without any internal bias.

A. A random sample of 1,200 people from a list of all U.S. residential addresses

Question 7

1 point

The median age of ten people in a room is 50 years. One 40-year-old person leaves the room. What can we expect will happen to the median age for the remaining nine people? Do we know the median age of those nine people; if so, what is it?

Ans 7:

We cannot estimate the median value , the median value shifts to the right

Question 8

1 point

The mean age of ten people in a room is 50 years. One 70-years-old person leaves the room. What can we expect will happen to the mean age for the remaining nine people? Do we know the mean age of those nine people; if so, what is it?

Ans: 8

Yes we do the mean will shift to the left. the mean is 45.5.

Question 9

1 point

Which of the following sets of numbers has the largest standard deviation?

A. 2, 4, 6, 8
B. 7, 8, 9, 10
C. 5, 5, 5, 5
D. 1, 2, 3, 5

```
sd_1 = sd(c(2, 4, 6, 8))
sd_2 = sd(c(7, 8, 9, 10))
sd_3 = sd(c(5, 5, 5, 5))
sd_4 = sd(c(1, 2, 3, 5))

max(sd_1, sd_2, sd_3, sd_4)
```

```
[1] 2.581989
```

The max value is option C

Question 10

6 points

I hypothesize that people with greater social trust are more likely to turnout to vote in American national elections. I use data from the 2012 General Social Survey to examine how respondents' self-reported level of social trust might be correlated with their decisions to vote. In particular, I use the `social_trust` variable (i.e., a 4-point ordinal indicator of social trust –

larger values reflect greater trust) and the `vote08` variable (i.e., a dichotomous indicator where a `1` indicates that the respondent voted) to test my hypothesis.

Complete the cross-tab below so that you may properly evaluate my hypothesis. Briefly interpret the results of your completed cross-tab. Do the data suggest that social trust is related to voting in 2008? Be sure to explain the nature of the relationship (or lack thereof, if relevant).

> **Note**
>
> Table entries represent raw counts of observations within each cell.

```
#Creating the table:
Cross_tab_data <- as.data.frame(matrix(data = NA, nrow = 4, ncol = 3))
names(Cross_tab_data) <- c('Social Trust', 'Vote08_1', 'Vote08_0')

Cross_tab_data[1,1] <- 0
Cross_tab_data[1,2] <- 257
Cross_tab_data[1,3] <- 137
Cross_tab_data[2,1] <- 1
Cross_tab_data[2,2] <- 194
Cross_tab_data[2,3] <- 93
Cross_tab_data[3,1] <- 2
Cross_tab_data[3,2] <- 192
Cross_tab_data[3,3] <- 56
Cross_tab_data[4,1] <- 3
Cross_tab_data[4,2] <- 240
Cross_tab_data[4,3] <- 30

Cross_tab_data
```

```
  Social Trust Vote08_1 Vote08_0
1            0      257      137
2            1      194       93
3            2      192       56
4            3      240       30
```

Based on the above data, we can see that the higher the social trust, the less likely people would would vote. However I do not see the effect to be continuous, even on lower trust, there was significant voting.

Applied Questions

Please include your R code. All data sets referenced below are available through the `poliscidata` R-package.

Total points: 20

Question 1

10 points total

Use the `states` dataset (the U.S. state is the unit of analysis) and estimate a bivariate regression where the size of a state's urban population (`urban`) explains variation in abortion attitudes (`permit`) and report the results in a professionally formatted table. The variable `permit` measures the percentage (on a 0-to-100 scale) of a state's population that says abortion should always be allowed. The variable `urban` measures the percentage (on a 0-to-100 scale) of a state's population in an urban area. Answer the following questions.

A. Interpret the effect of the independent variable on the dependent variable. *2 points*

B. Interpret the estimate of the intercept. Is it substantively meaningful to interpret this coefficient on its own? Explain why, or why not. *2 points*

C. Compute the residual sum of squares for the following two observations combined: (1) California; and (2) Texas. *2 points*

D. How well does the model fit the data (i.e., how well can we explain abortion attitudes with this model?) *2 points*

E. Is the relationship between the independent and dependent variable causal? Why or why not? *2 points*

Ans 1:

```
States_data <- poliscidata::states
```

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

```
#Creating the linear model
states_linear_model <- lm(permit ~ urban, data = States_data)
modelsummary::modelsummary(states_linear_model)
```

	(1)
(Intercept)	9.639
	(7.341)
urban	0.373
	(0.100)
Num.Obs.	40
R2	0.270
R2 Adj.	0.251
AIC	289.8
BIC	294.8
Log.Lik.	−141.890
RMSE	8.40

The variable 'urban' has a positive coefficient as well as being statistically significant. This means that the variable has a significant and positive effect on abortion attitudes. The Rsquared and the Adjusted Rsquared values are not very high however, this means that there is significant variance that is not being explained.

2. The intercept term is the predicted value of the dependent variable when all the independent terms are equal to zero. In this case it is 9.639, that means that when the urban value is zero, this will be the percentage that says that abortion should be allowed. Intercepts should generally be carefully evaluated in regression situations as they are just a mathematical optimization. In this case, it seems alright to interpret the variable, as it indicates the abortion acceptance rate in a completely rural state.

3. Calculating the value for Texas and California:

```
States_data <- as.data.table(States_data)

data_cal <- States_data[stateid == as.character(unique(States_data$stateid)[5]),]

#Extracting the data
Predicted_cal <- predict(states_linear_model, data = data_cal$urban)[6] #Predicting u
Residual_cal = data_cal$permit - Predicted_cal

data_tx <- States_data[stateid == as.character(unique(States_data$stateid)[43]),]

#Extracting the data
Predicted_tx <- predict(states_linear_model, data = data_tx$urban)[6]

#Predicting using lm model
```

```
Residual_tx = data_tx$permit - Predicted_tx

sum(Residual_cal^2, Residual_tx^2)
```

```
[1] 210.1869
```

4. The Rsquared value is 0.27, this means that hardly 27% of the variance was explained by the model. Although this is acceptable in some cases, this would generally not be considered a good model.

5. In this case, the relationship is not causal. Although the correlation effects between the two variables are positive, this is not enough to prove causality. The lack of significant rsquare is a major problem, to prove causality we would expect the variance to be a little higher. Even then, statistical tests such as this are not enough to prove causality. It is best proven in an observational set up, where multiple draws of the data can be performed to prove the effect. In this particular case, we can understand that an urban population can imbibe values that increase support for abortion, but the urbanization of a state is not directly causing the increase in support for abortion.

Question 2

5 points

Use the `gss` data set (the unit of analysis is the individual survey respondent) and evaluate the hypothesis that Republicans had less confidence in the executive branch of the federal government than Democrats in 2016. Use the following variables: `partyid` is a 7-category ordinal indicator (0 = Strong Democrat; 1 = Weak Democrat; 2 = Independent Democrat; 3 = Independent; 4 = Independent Republican; 5 = Weak Republican; 6 = Strong Republican); and `confed` is a 3-category ordinal indicator (1 = "a great deal" of confidence; 2 = "only some;" 3 = "hardly any"). Do the data support the hypothesis and how do you know?

Ans 2:

```
#We can construct a cross tab to measure the effects of this variable
gss_data <- poliscidata::gss

gss_data <- poliscidata::gss

gss_data$partyid <- as.character(droplevels(gss_data$partyid))

gss_data['partyid'][gss_data['partyid'] == 'StrDem'] <- '0'
gss_data['partyid'][gss_data['partyid'] == 'WkDem'] <- '1'
gss_data['partyid'][gss_data['partyid'] == 'IndDem'] <- '2'
```

confed		0	1	2	3	4	5	6	All
1	N	71	35	31	28	8	14	6	196
	% row	36.2	17.9	15.8	14.3	4.1	7.1	3.1	100.0
2	N	110	135	91	113	36	75	40	618
	% row	17.8	21.8	14.7	18.3	5.8	12.1	6.5	100.0
3	N	53	59	44	91	59	81	79	493
	% row	10.8	12.0	8.9	18.5	12.0	16.4	16.0	100.0
All	N	356	343	235	373	157	250	192	1974
	% row	18.0	17.4	11.9	18.9	8.0	12.7	9.7	100.0

```
gss_data['partyid'][gss_data['partyid'] == 'Ind'] <- '3'
gss_data['partyid'][gss_data['partyid'] == 'IndRep'] <- '4'
gss_data['partyid'][gss_data['partyid'] == 'WkRep'] <- '5'
gss_data['partyid'][gss_data['partyid'] == 'StrRep'] <- '6'

gss_data$confed <- as.character(droplevels(gss_data$confed))

gss_data['confed'][gss_data['confed'] == 'A GREAT DEAL'] <- '1'
gss_data['confed'][gss_data['confed'] == 'ONLY SOME'] <- '2'
gss_data['confed'][gss_data['confed'] == 'HARDLY ANY'] <- '3'


datasummary_crosstab(confed ~ partyid, data = gss_data)
```

Based on the crosstab evaluation, we can see that the combination of 0,1 and 2 categories in having a great deal of confidence is 70% of the total number of rows. Republicans had about 44% of the rows where they had hardly any confidence in the government. Republicans having high confidence in the government is only 15% of the total number of rows as well.

Based on these percentage values, we can say with confidence that democrats had a higher trust in the executive government compared to republicans in 2012.

Question 3

5 points

Use the `world` dataset and evaluate the distributions for each the following variables: `literacy` (a country's literacy rate) and `free_overall` (a country's degree of economic freedom). Be sure to visually display each distribution and thoroughly describe their key attributes. Next, evaluate the bivariate relationship between the two variables – i.e., is economic freedom associated with literacy? If so, what is the nature of the relationship and how do you know? In

doing so, be sure to use proper descriptive tools (and thus do not rely simply on a regression output).

Ans 3:

```
world_data <- poliscidata::world

#Summary
summary(poliscidata::world$literacy)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  21.80   67.90   88.70   80.61   98.20  100.00      10
```

```
length(poliscidata::world$literacy)
```

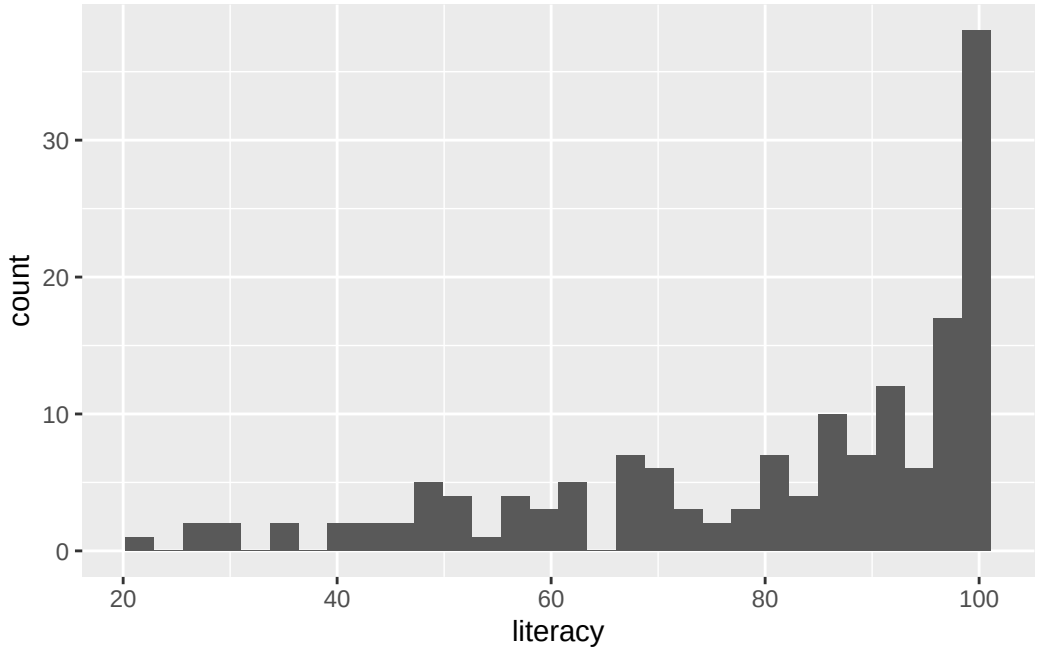
```
[1] 167
```

Taking a look at the summary statistics, we can see that the data contains 10 NA values out of a total sample size of 167. It has a median percentage value of 88.7% with a max of 100% and a minimum of 21%. As this is a percentage scale, the data can range from 0 to 100. From this data alone we can see that the distribution skews to the right.

Now let's see the distributions

```
#Histogram chart
ggplot(data = poliscidata::world, aes(x=literacy)) + geom_histogram(na.rm = TRUE)
```

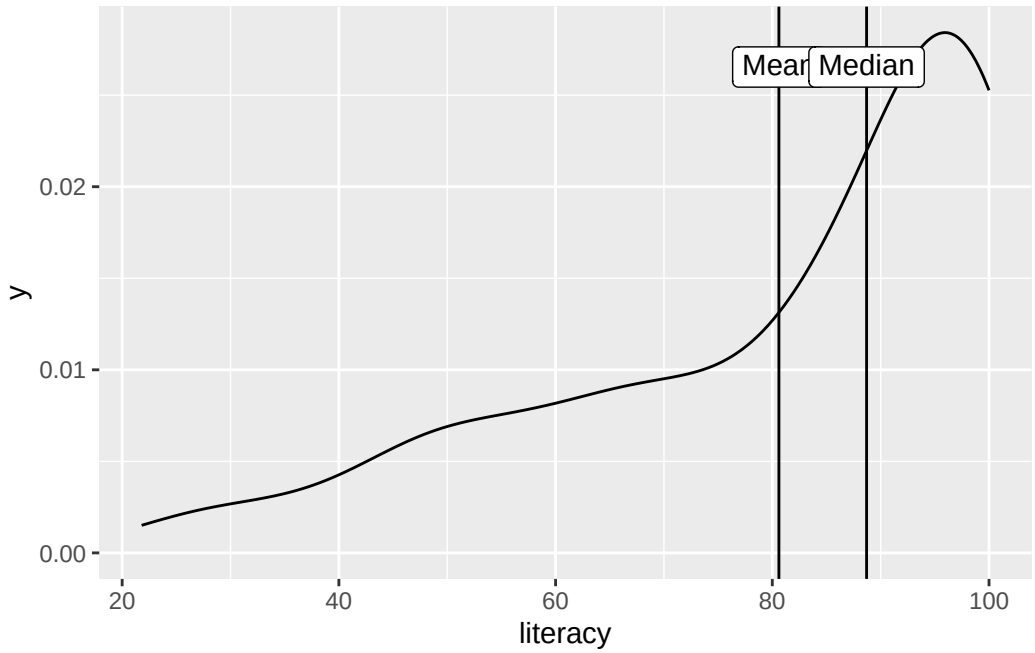
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Plotting the data series and removing the NA values, we can see that our original hypothesis was correct with a majority of the data above the 80% mark.

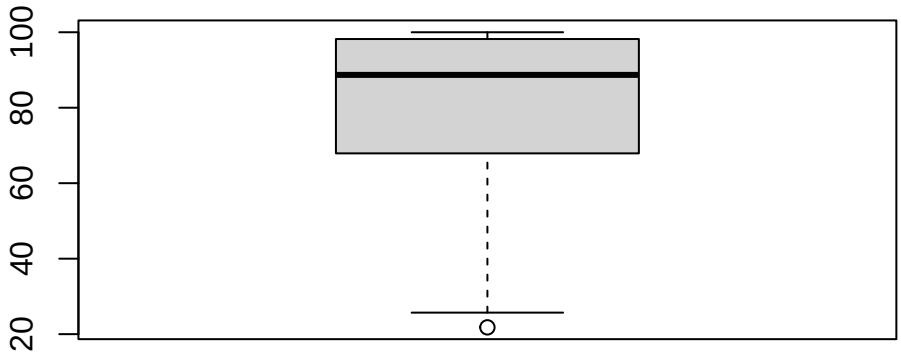
Taking a look at the density chart,

```
ggplot(data = poliscidata::world, aes(x=literacy)) +
  geom_density(na.rm = TRUE) +
  geom_vline(xintercept=mean(poliscidata::world$literacy, na.rm = TRUE)) +
  geom_vline(xintercept=median(poliscidata::world$literacy, na.rm = TRUE)) +
  annotate(x = mean(poliscidata::world$literacy, na.rm = TRUE), y = +Inf, label = "Mean",
  annotate(x = median(poliscidata::world$literacy, na.rm = TRUE), y = +Inf, label = "Media
```



Looking at the density chart we can clearly see the majority of the data skewing to the right and above 80%.

```
boxplot(poliscidata::world$literacy, na.rm = TRUE)
```



Doing a quick outlier analysis as well, the data is definitely skewed, with the singular outlier lying below the series.

Let's now explore the same for the free_overall variable

```
#Summary
summary(poliscidata::world$free_overall)
```

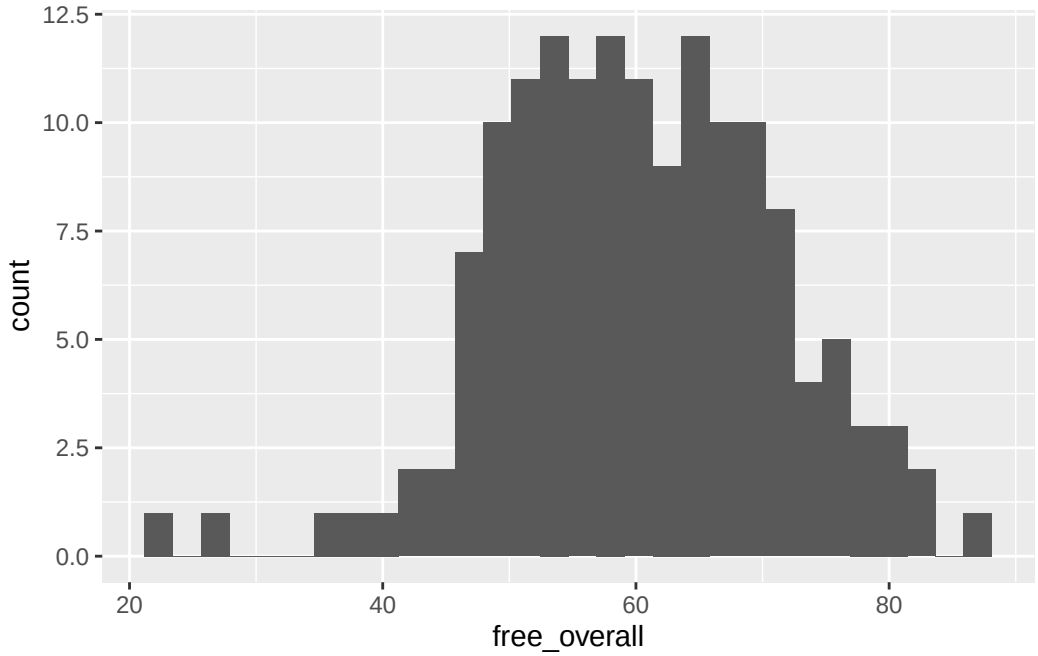
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  21.40   52.55   59.45   60.04   67.70   86.10      17
```

Taking a look at the summary statistics, we can see that the data contains 17 NA values out of a total sample size of 167. It has a median percentage value of 59.45% with a max of 86.10% and a minimum of 21.40%. As this data series is again a percentage distribution, the data can range from 0% to 100%. From this data alone we can see that the distribution is a lot more normal than the distribution for literacy.

Let's see the distributions:

```
#Histogram chart
ggplot(data = poliscidata::world, aes(x=free_overall)) + geom_histogram(na.rm = TRUE)
```

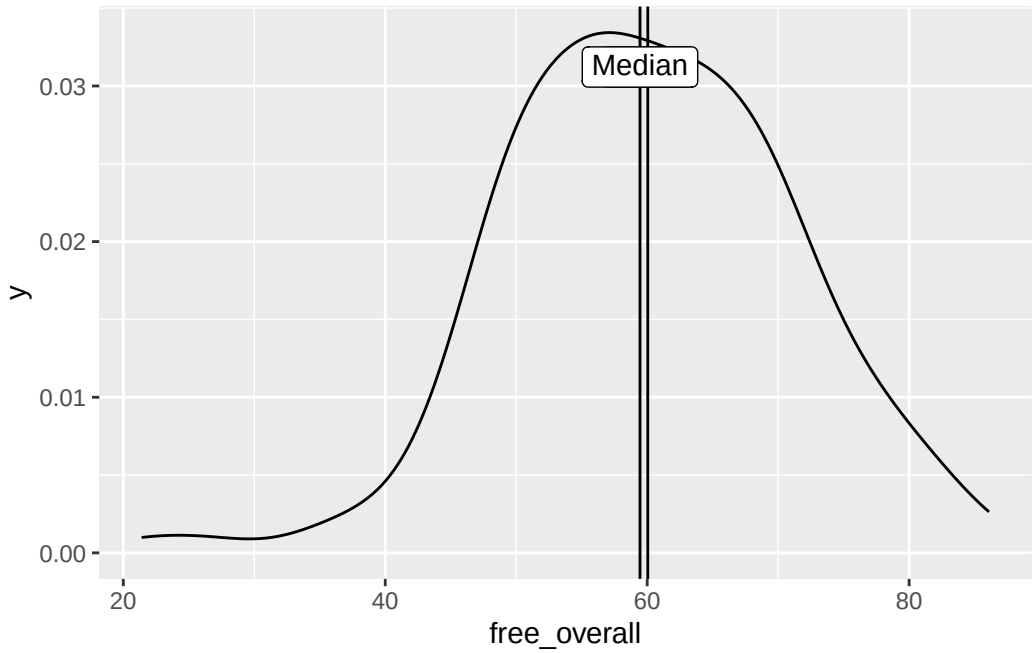
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see that most of the data points are clearly clustered in the center of the distribution.

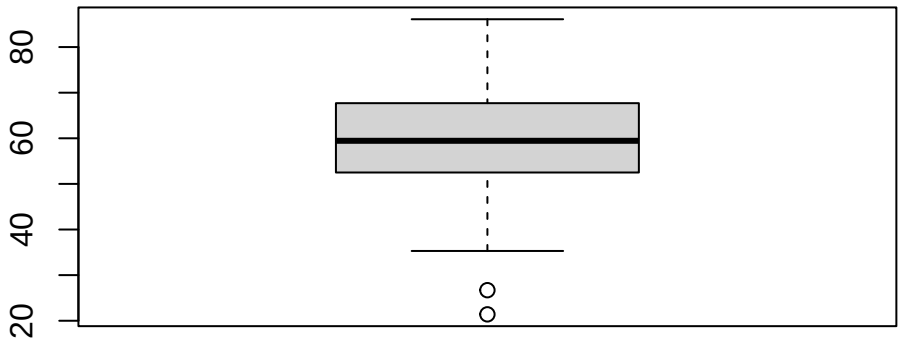
Let's see a density plot

```
ggplot(data = poliscidata::world, aes(x=free_overall)) +
  geom_density(na.rm = TRUE) +
  geom_vline(xintercept=mean(poliscidata::world$free_overall, na.rm = TRUE)) +
  geom_vline(xintercept=median(poliscidata::world$free_overall, na.rm = TRUE)) +
  annotate(x = mean(poliscidata::world$free_overall, na.rm = TRUE), y = +Inf, label = "Mea
  annotate(x = median(poliscidata::world$free_overall, na.rm = TRUE), y = +Inf, label = "M
```



The mean and the median are very close to each other and almost touch the peak of the curve. The data is very evenly distributed in this scenario.

```
boxplot(poliscidata::world$free_overall, na.rm = TRUE)
```

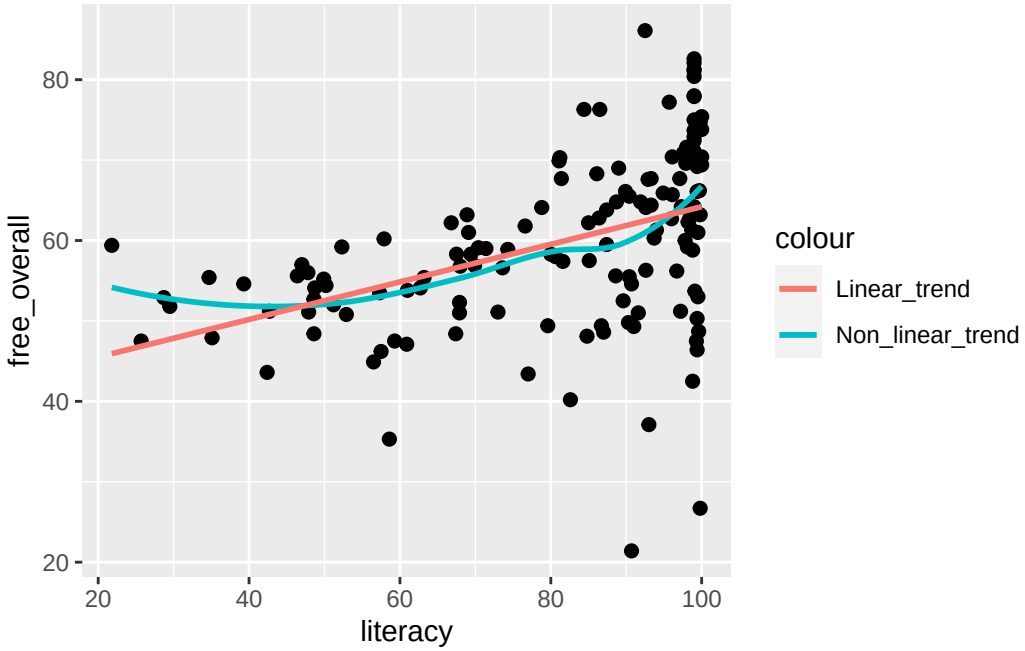


Compared to the literacy variable, the box plot looks at lot more centered. There are however two outliers in this case rather than the one before.

Examining the relationship between the to variables we can plot the two variables and see the interacting trend between them to get a rough idea of the relationship.

```
ggplot(data = world_data, aes(x=literacy, y=free_overall)) +
  geom_point(size=2, na.rm = TRUE) +
  geom_smooth(method=loess, se=FALSE, na.rm = TRUE, aes(colour="Non_linear_trend")) +
  geom_smooth(method=lm, se=FALSE, na.rm = TRUE, aes(colour="Linear_trend"))
```

```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```



Based on the plot, there seems to be a rather weak positive relationship between the variables.

Let's check the linear correlation level,

```
#Extracting the columns needed dropping the NA values and checking the linear correlation
world_literacy_free_data <- as.data.frame(cbind(world_data$literacy, world_data$free_overa
world_literacy_free_data <- na.omit(world_literacy_free_data)
row.names(world_literacy_free_data) <- NULL

cor(world_literacy_free_data)[1,2]
```

```
[1] 0.4332167
```

As we can see there is a 0.43 level of correlation between the two variables. This is in the positive direction, but maybe not very strong.

Let's now see if we can explain the variance of one variable with the another variable.

First let's test if we can explain the variance of literacy based on overall freedom

	(1)
(Intercept)	33.078
	(8.552)
free_overall	0.803
	(0.141)
Num.Obs.	143
R2	0.188
R2 Adj.	0.182
AIC	1243.2
BIC	1252.1
Log.Lik.	−618.613
RMSE	18.30

	(1)
(Intercept)	40.831
	(3.422)
literacy	0.234
	(0.041)
Num.Obs.	143
R2	0.188
R2 Adj.	0.182
AIC	1066.7
BIC	1075.6
Log.Lik.	−530.332
RMSE	9.87

```
lm_obj_literacy_v_freedom <- lm(literacy ~ free_overall, data = world_data)
modelsummary::modelsummary(lm_obj_literacy_v_freedom)
```

Based on the model summary, we can see that the free_overall variable is positive and statistically significant in explaining the variation of literacy. This might suggest that the more economically free a populace is, the more literate they can become. We can see both R squared and the Adj, R square to be reasonably low indicating that there are probably many more reasons affecting the variance of literacy.

How about the other way around ,

```
lm_obj_freedom_v_literacy <- lm(free_overall ~ literacy, data = world_data)
modelsummary::modelsummary(lm_obj_freedom_v_literacy)
```

Based on the model summary, we can see that the literacy variable is positive and statistically significant in explaining the variation of economic freedom. This might suggest that the more literate a populace is, the more economically free they they can become. We can see both R squared and the Adj, R square to be reasonably low indicating that there are probably many more reasons affecting the variance.

But based on all our experimentation, we can conclude that there is a weak but significant effect of the two variables on each other.