HW 4: Financial Stressors and Cognitive Performance

Do changes in one's financial circumstances affect one's decision-making process and cognitive capacity? In a research study, researchers randomly selected a group of US respondents to be surveyed before their payday and another group to be surveyed after their payday. People working for different employers or in different jobs get paid at different times. So at any given point in time, some people will have just been paid and others will be waiting to be paid. Under this design, the researchers believed that the respondents of the `Before Payday` group would be more likely to be financially strained (which we will consider to be the treatment) than those of the `After Payday` group (which we will consider to be the alternative). The researchers were interested in investigating whether or not being financially strained affects people's decision making and cognitive performance. Other researchers have found that in a lab setting, scarcity induces an additional mental load that impedes cognitive capacity.

In this study, the researchers administered a number of decision-making and cognitive performance tasks to the `Before Payday` and `After Payday` groups. We focus on the *Numerical Stroop Task*, which measures cognitive control. An example of a Numerical Stroop Task would be to present two numbers side by side, and vary the font sizes so that sometimes the lower number is in a bigger font than the higher number and vice versa and test how long it takes the subject to correctly identify which number is larger and how many errors they make. In general, taking more time to complete this task indicates less cognitive control and reduced cognitive ability. They also measured the amount of cash the respondents have on hand, the amount of money in their checking and saving accounts, and whether or not they have a low annual income. The data set is in the CSV file `poverty.csv`. The names and descriptions of variables are given below:

Name	Description
`treatment`	Treatment conditions: `Before Payday` and `After Payday`
`cash`	Amount of cash respondent has on hand, in dollars
`accts_amt`	Amount in checking and saving accounts, in dollars
`stroop_time`	Log-transformed average response time for cognitive stroop test
`income_less20k`	Binary variable: `1` if respondent earns less than 20k a year and `0` otherwise

Question 1 (7 points)

1a (2 points)

What is the specific causal question the researchers would like to answer?

1b (2 points)

For the subjects who had not yet received their paycheck, what is the average missing counterfactual at the group level?

1c (3 points)

What strategy does this study propose using to estimate this missing counterfactual? What assumption is required for this estimate of the missing counterfactual to be unbiased?

Answer 1

Your Text answers for 1a - 1c go below:

1a

What is the impact of the financial situation before payday , relative to the financial situation after payday , on cognitive capacity and decision making for U.S respondants that took part in this survey/study? ### 1b

not received pay : AVG MCF : What the average cognitive capacity/decision making skills would have been for all the subjects that did not receive their paycheck on payday if instead they received their paychecks but all else remained the same.* ### 1c

we can use the post only cross sectional study to estimate the missing counter factual. the assumption that is required to estimate the MCF is that there should be no systematic differences in observed or unobserved baseline covariates between the Subjects that receive their pay check and the subjects that don't receive their paycheck and that that if there were any differences they must not be associated with the outcome that is the "average cognitive capacity" as this could bias the MCF.

Question 2 (10 points)

Load the `poverty.csv` data set. Use histograms and boxplots to examine the univariate distributions of the two baseline covariates that are financial resources measures: `cash` and `accts_amt`(include all observations/rows in the data set in each figure). Calculate the mean and standard deviation, median, quartiles, and IQR for the two financial resources measures for all subjects in the dataset (not separately by treatment group).

Describe these two variables' univariate distributions based on the figures and the summary statistics.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     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 become errors
```

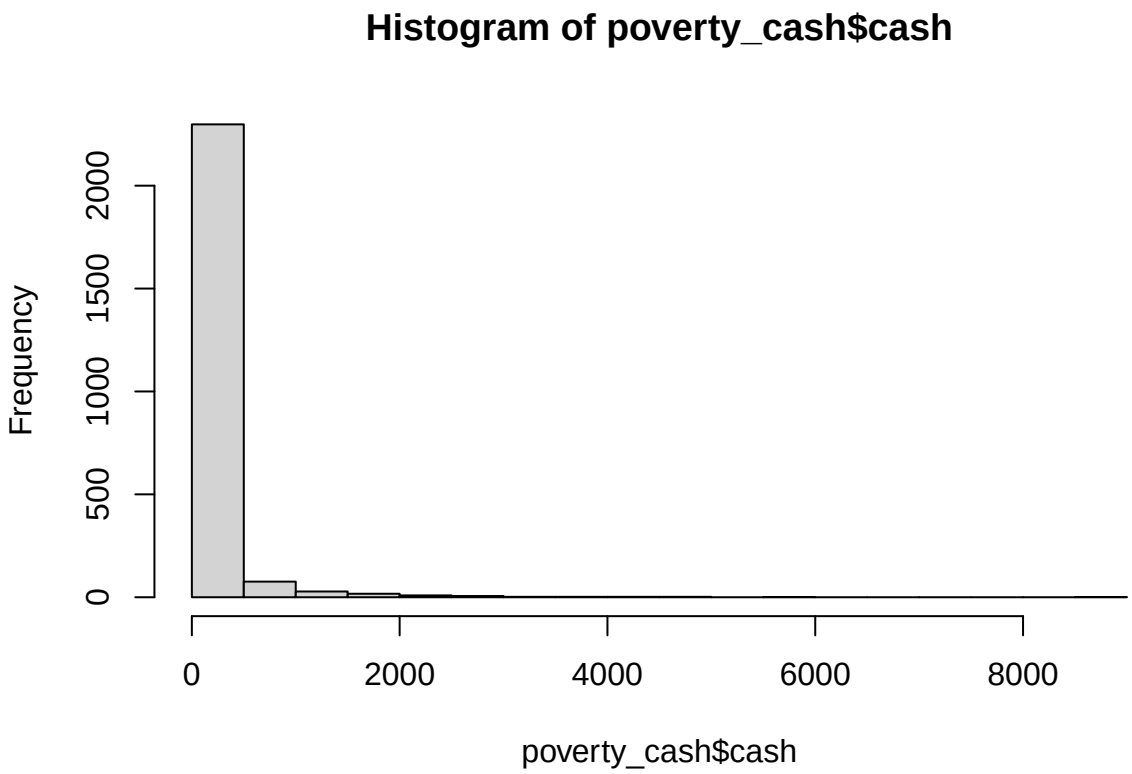
```
library(ggplot2)

# Load the poverty data; remember to put the csv file in a data sub-folder!
 poverty <- read.csv("data0/poverty.csv")
 # Log-transform money in cash
 poverty$log_cash <- poverty$cash
 poverty$log_cash[poverty$log_cash == 0] <- 1
poverty$log_cash <- log(poverty$log_cash)
```

Answer 2

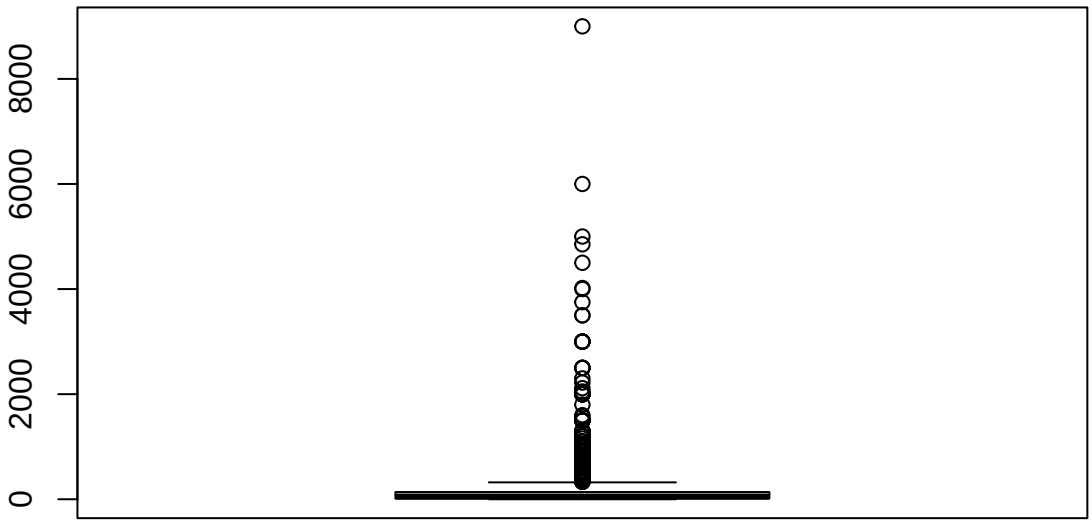
Your answers for Question 2 go here!

2 Cash Variable

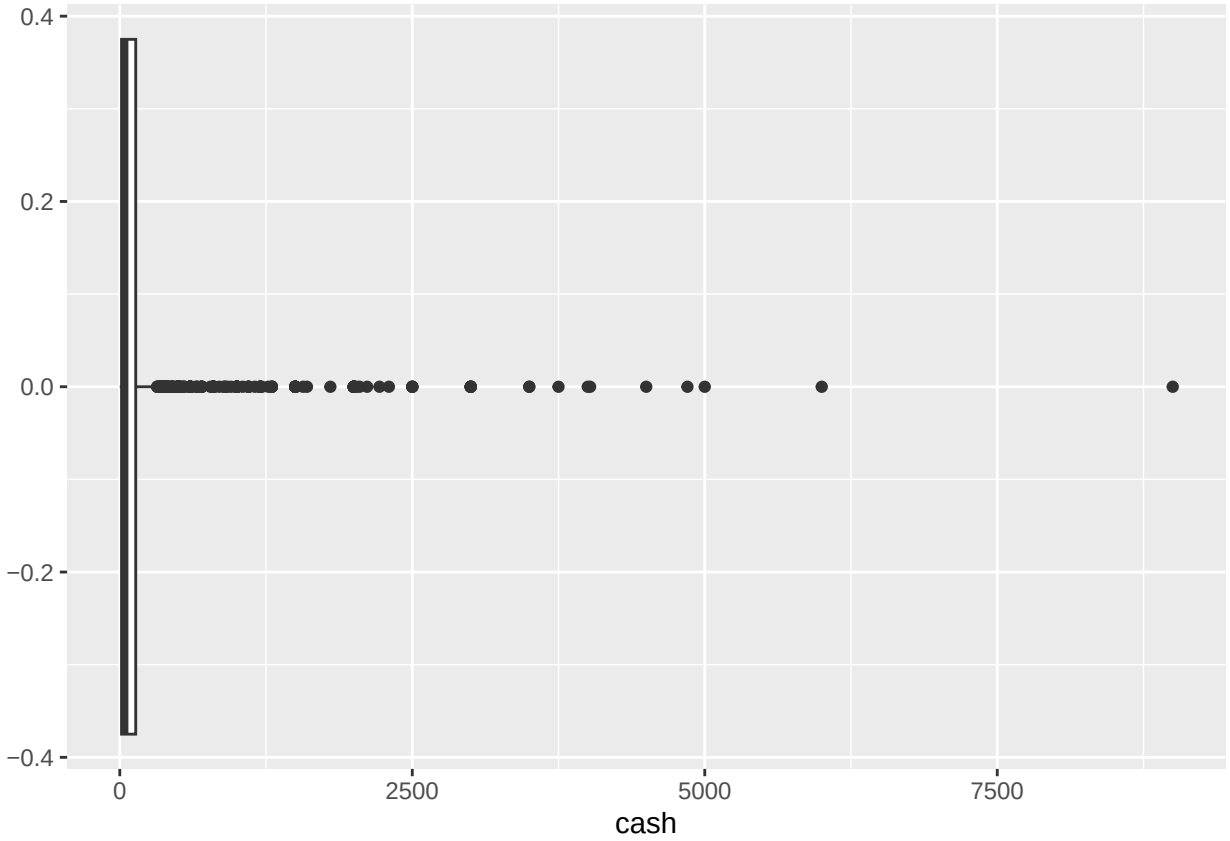
```
# Your code for the cash variable goes here
poverty_cash <- poverty[!is.na(poverty$cash),]
hist(poverty_cash$cash)
```




```
boxplot(poverty$cash)
boxplot(poverty_cash$cash)
```




```
mean_cash <- mean(poverty$cash, na.rm = TRUE)
poverty_cash %>%
  ggplot(aes(x = cash)) +
  geom_boxplot()
```




```
summary(poverty$cash, na.rm = TRUE )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    15.0    49.5   169.0   136.2  9000.0     226
```

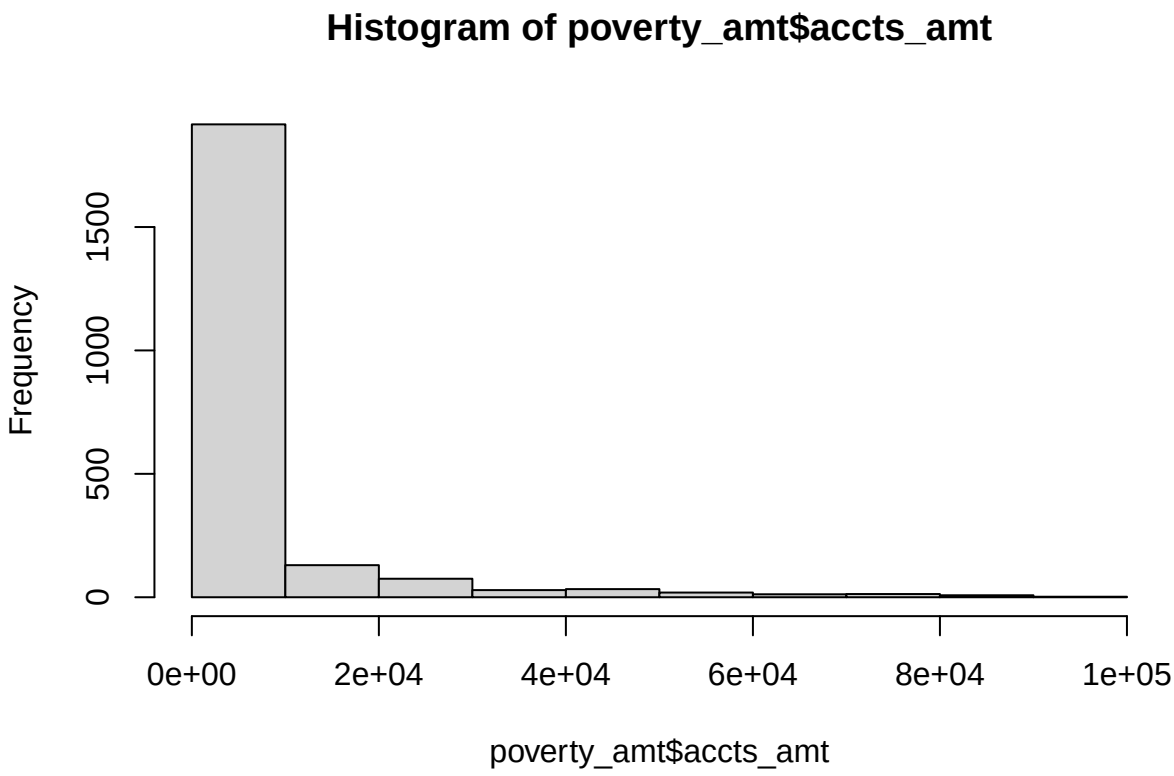
```
sd(poverty_cash$cash)
```

```
## [1] 451.283
```

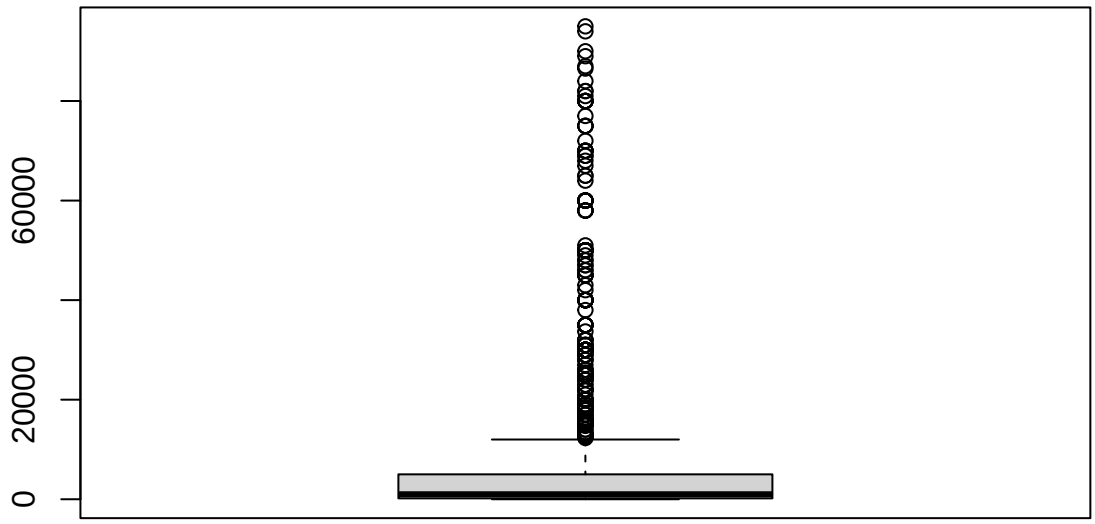
2 Cash Variable Text Answer:

the mean is 169. The median is 49.5, the 1st and 3rd quartiles are 15 and 136.2 respectively.the IQR is 136.2 - 15 = 121.2 range is between 0- 9000.50 percent of the values lie between 15 and 136.2. the sd is 451.28. it is unimodal with a modal value of at O. There are no obervable gaps in the distribution. Almost all of the values lie between 0 and 2500 with a right tail right skew. No spikes are visible. however with the boxplot we can observe that Most of the outliers are connected to the upper quartile. between 2300 - 9000 there are several outliers that cannot be observed in the histogram.

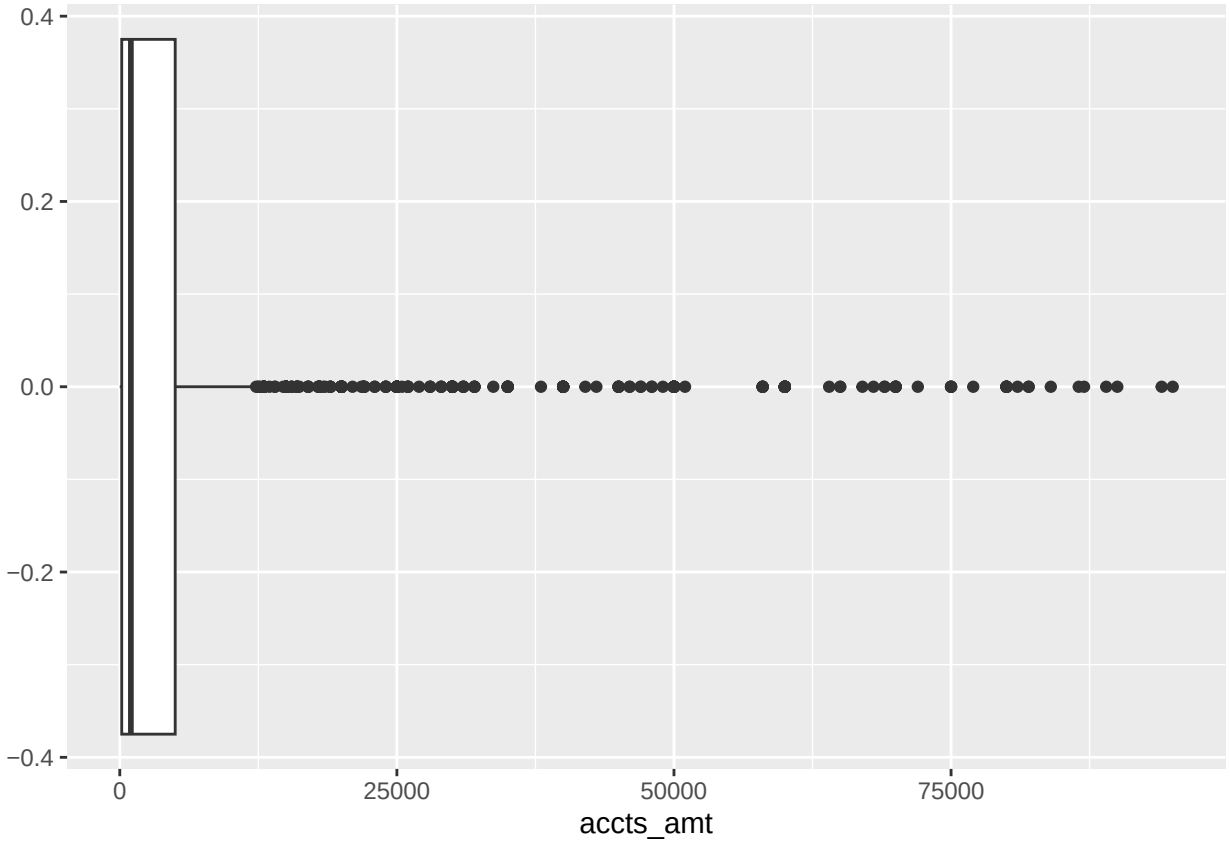
```
poverty_amt <- poverty[!is.na(poverty$accts_amt),]
hist(poverty_amt$accts_amt)
```




```
boxplot(poverty$accts_amt)
```




```
mean_acc <- mean(poverty$accts_amt, na.rm = TRUE)
poverty_amt%>%
  ggplot(aes(x = accts_amt)) +
  geom_boxplot()
```




```
summary(poverty$accts_amt , na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     176    1000    6211    5000   95000     433
```

```
sd(poverty_amt$accts_amt)
```

```
## [1] 13517.69
```

2 Accounts Variable Text Answer:

the mean is 6211. The median is 1000, the 1st and 3rd quartiles are 176 and 5000. respectively.The IQR is 5000 - 176 = 4824. The range is between 0- 95000.50 percent of the values lie between 176 and 5000. the sd is 13517.69 it is unimodal with a modal value of at O. There are no obervable gaps in the distribution. It is right skewed. No spikes are visible. however with the boxplot we can observe that Most of the outliers are connected to the upper quartile and range until 51000. between 51000 - 90000 there are several outliers that cannot be observed in the histo between 1/2 or 1/3 Iqrs apart from each other.

Question 3 (9 points)

3a (7 points)

Now, use the code provided below to take the *natural logarithm* of the `cash` variable and summarize the transformed variable using summary statistics and figures: calculate the mean, median, standard deviation, and IQR of the log transformed `cash` variable (again for all observations in the dataset). Create a boxplot and histogram for the transformed variable. Describe the log transformed variable's distribution based on the figures and the summary statistics. How does this distribution differ from the `cash` variable on the original scale?

3b (2 points)

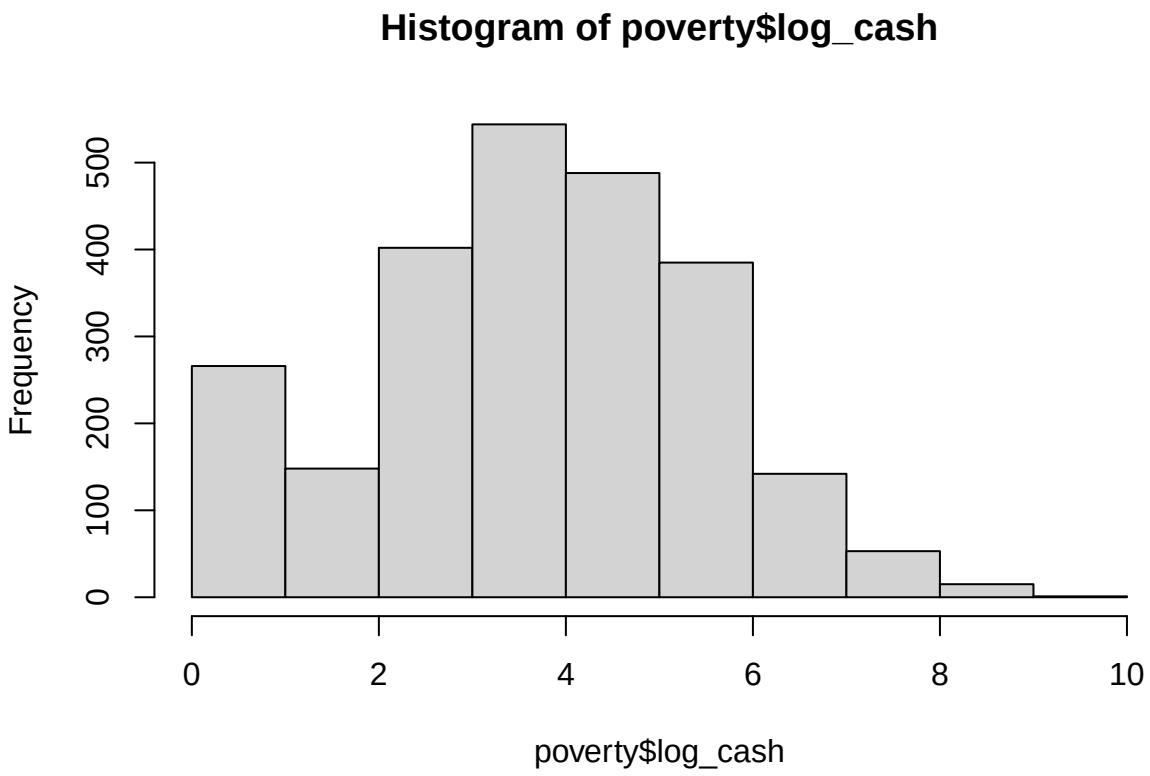
State an advantage and a disadvantage of transforming the data in this way.

NOTE FOR INTERESTED STUDENTS: *Since the natural logarithm of 0 is undefined, researchers often add a small value (in this case, we will use \$1 so that* $\log 1 = 0$*) to the 0 values for the variables being transformed (in this case,* `cash` *and* `accts_amt`*) in order to successfully apply the* `log()` *function to all values. Note that we do this recoding only for the purposes of taking the logarithmic transformation – we have kept the original variables the same.*

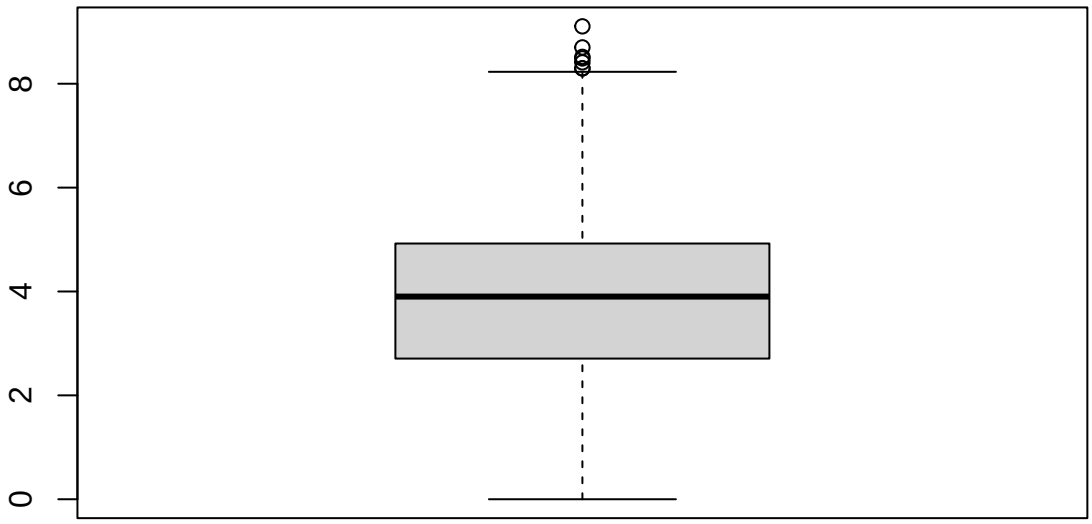
```
### Log transformation code for 3a
# Feel free to move this code where you need it

# Log-transform money in cash
 poverty$log_cash <- poverty$cash
 poverty$log_cash[poverty$log_cash == 0] <- 1
poverty$log_cash <- log(poverty$log_cash)
```

3a

```
# Your code for 3a goes here
hist(poverty$log_cash)
```

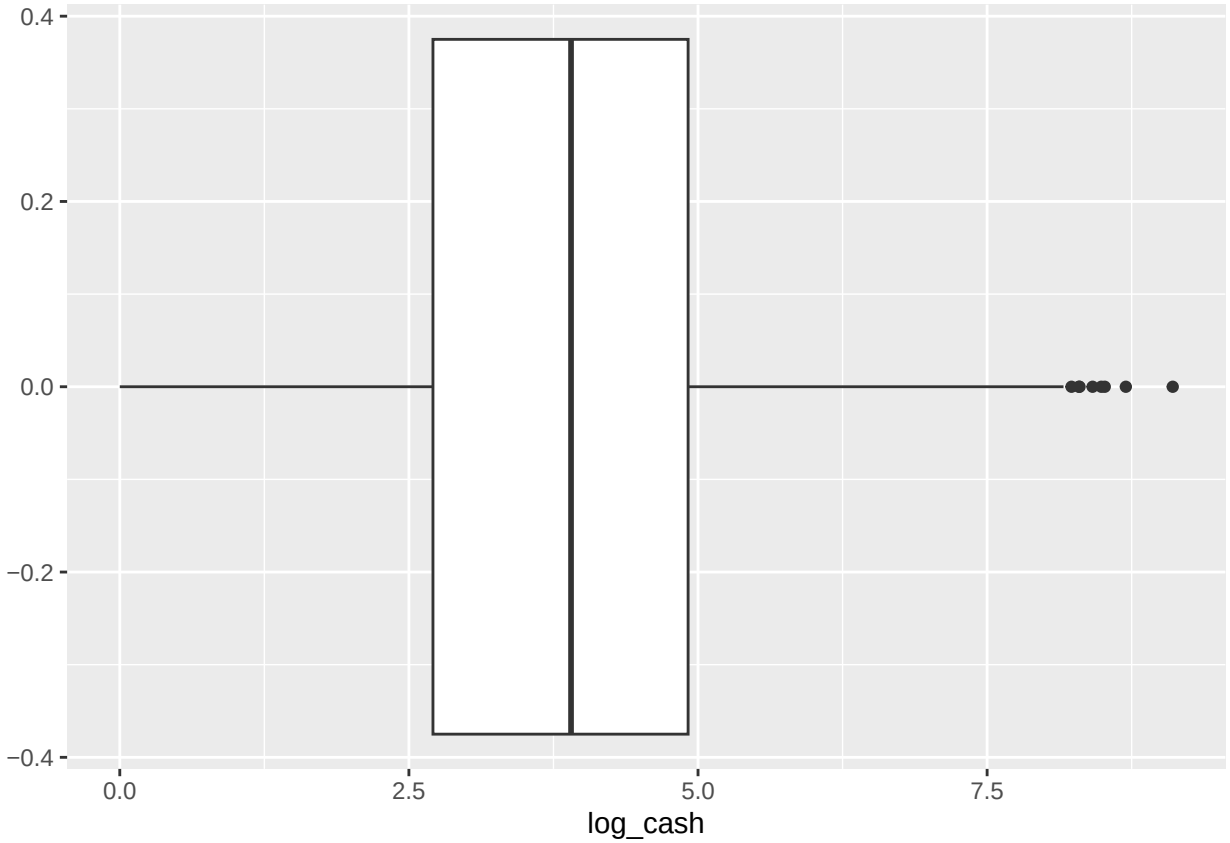



```
boxplot(poverty$log_cash)
```




```
mean_log_cash <- mean(poverty$log_cash, na.rm = TRUE)
poverty %>%
  ggplot(aes(x = log_cash)) +
  geom_boxplot()
```

```
## Warning: Removed 226 rows containing non-finite values (`stat_boxplot()`).
```




```
summary(poverty$log_cash, na.rm = TRUE )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.708   3.902   3.684   4.914   9.105     226
```

```
sd(poverty$log_cash , na.rm = TRUE)
```

```
## [1] 1.842889
```

Text Answer for 3a:

3b Text Answer:

the 1st quartile is 2.708. the third quartile 4.19. The median is 3.90. the mean is 3.68. the range is 0 - 9.105. the standard deviation is 1.84. the distribution is uni modal and skewed slightly to the right.it is slightly symmetric. 50 percent of the data is distributed between 2.7 and 4.91. the box plot is symmetric. the outliers are connected to the tail end with one outlier that is 1.5 iqr's away from the 3 rd quartile. the scale for the cash variable was in the e scale between oE - 1E+05 and the log scale was between 0-10 on the histo gram. the hsotogram for the cash variable seemed to have one modal with a strong right skew compared to the historgam produced for the log_cash variable with one modal and not a very storng right skew. the dsitubution for cash is not symmteric while symmetric for the log_cash.

Question 4 (9 points)

4a (5 points)

Now, let's focus on the primary outcome of interest for this study - cognitive performance. Let's estimate the effect of a presumed change in financial situation (in this case, waiting to get paid relative to just having been paid) on cognitive performance. Estimate the treatment's effect on the `stroop_time` variable (a log-transformed variable of the average response time for the Stroop Cognitive Test), using first the mean and then the median. What do these results tell you about the answer to the causal question from Q1?

4b (4 points)

Now compare the time it took subjects to complete the cognitive test across the `Before Payday` and `After Payday` groups using overlaid or side-by-side histograms and side-by-side boxplots of the `stroop_time`variable for the before and after payday groups. Based on these plots, do you think there is a meaningful difference in the time it took participants from each group to complete the cognitive test?

Answer 4

4a

```
poverty_paid <- poverty %>%
      filter(treatment == "After Payday")
poverty_unpaid <- poverty %>%
      filter(treatment == "Before Payday")
mean(poverty_paid$stroop_time, na.rm = TRUE)
```

```
## [1] 7.550519
```

```
median(poverty_paid$stroop_time, na.rm = TRUE)
```

```
## [1] 7.571088
```

```
mean(poverty_unpaid$stroop_time, na.rm = TRUE)
```

```
## [1] 7.539096
```

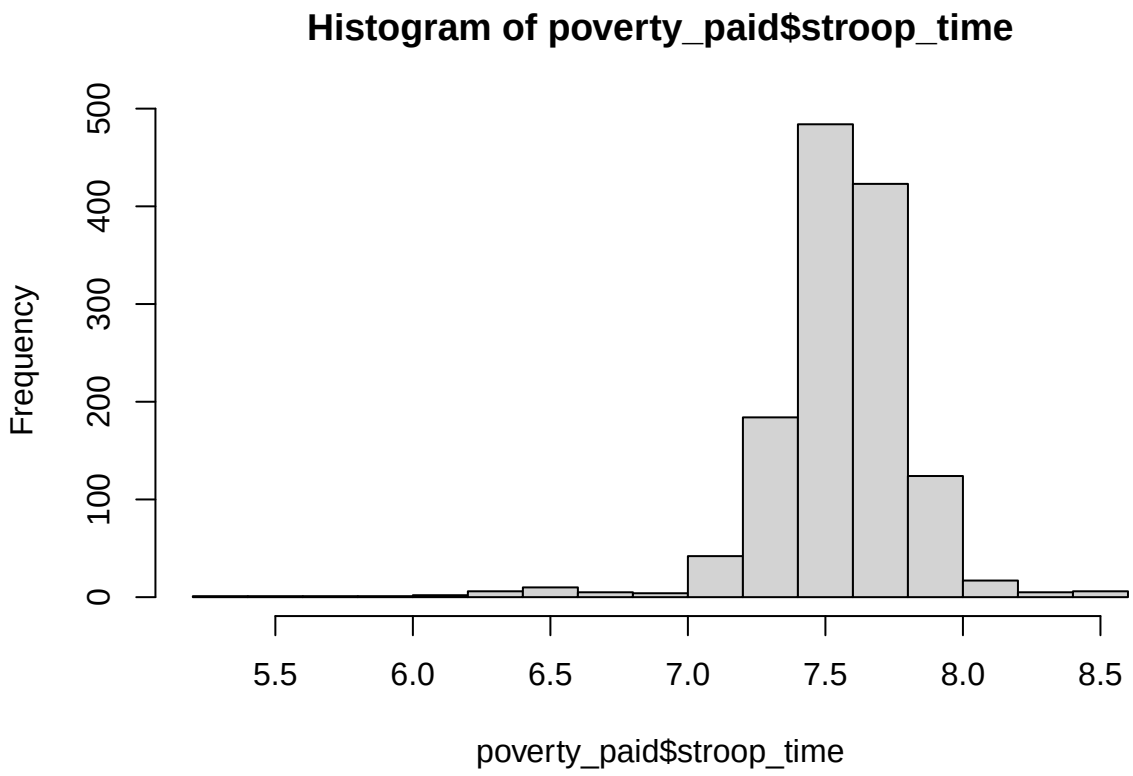
```
median(poverty_unpaid$stroop_time, na.rm = TRUE)
```

```
## [1] 7.556783
```

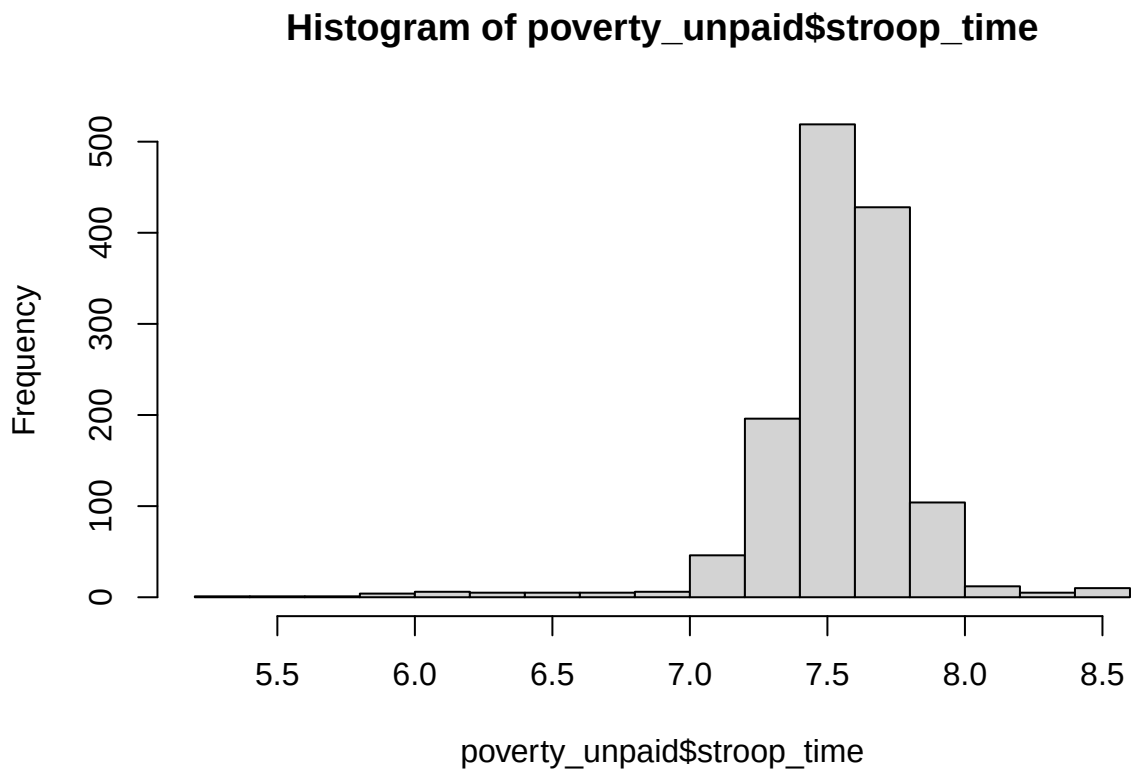
Text Answer for 4a:

a. the mean and median for the subjects that received their pay: 7.55 and 7.57
b. the mean and median for the subjects that didn't receive their pay: 7.539 and 7.556 the treatment effect might not significantly differ from the usual as the stroop time variable is approximately the same for both paid(alternative) and the not paid(treatment).therefore the impact on the cognitive ability do not differ significantly due to the financial stress upon not receiving the salary. ### 4b

```r
hist(poverty_paid$stroop_time)
```

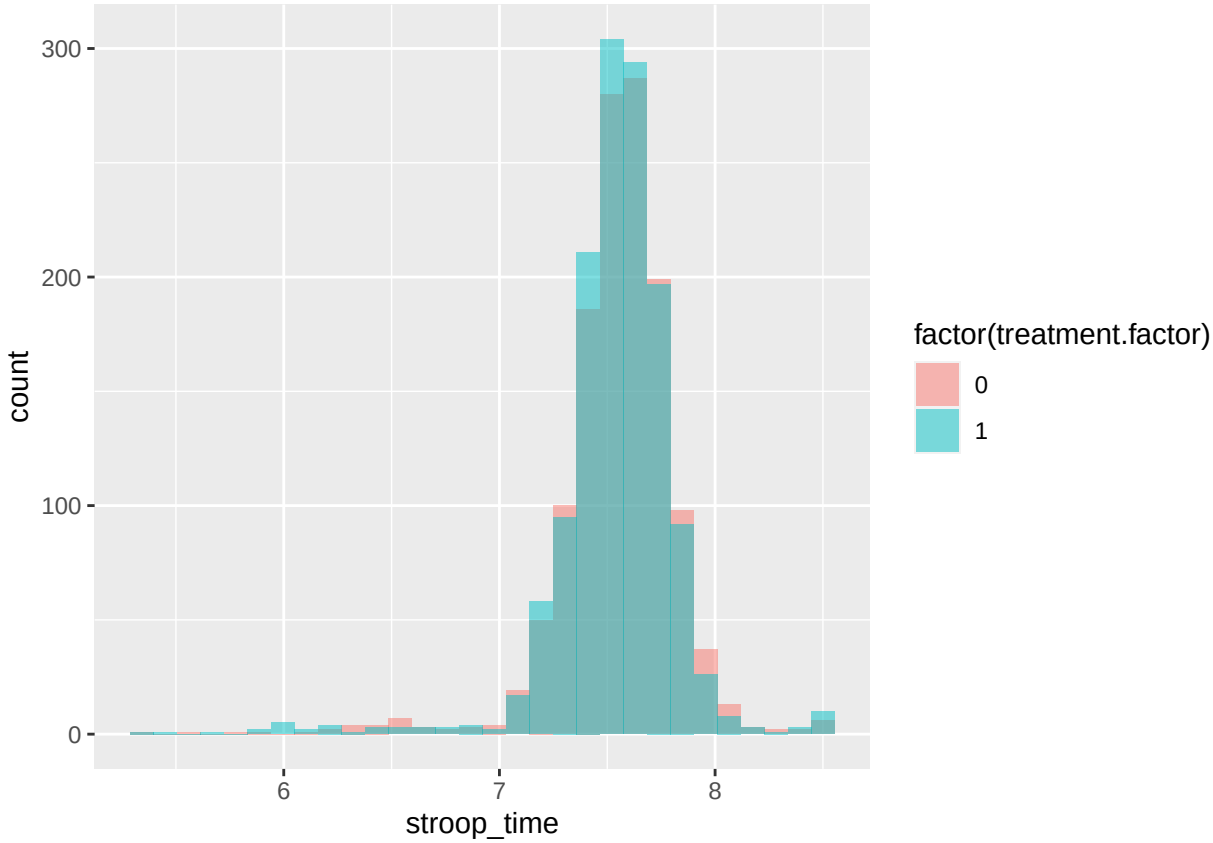



```r
hist(poverty_unpaid$stroop_time)
```

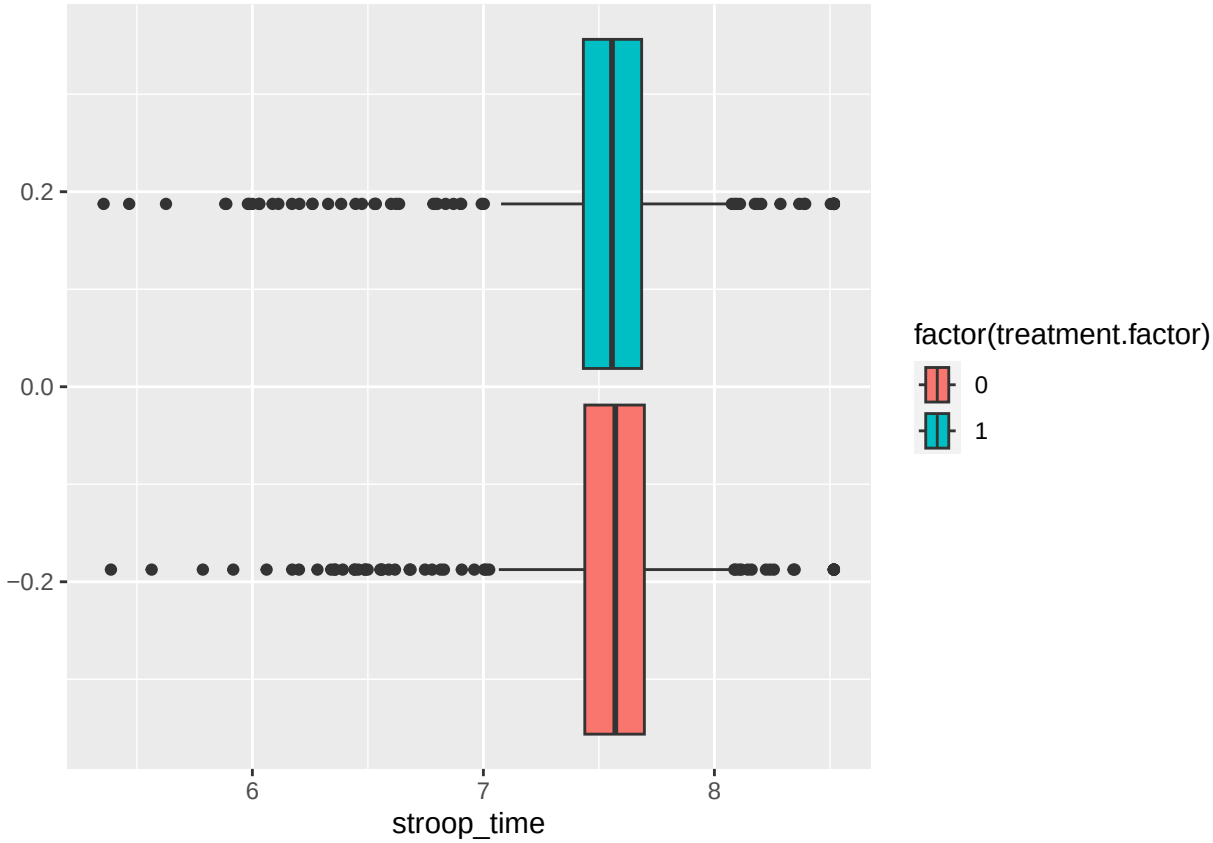



```
par(mfrow = c(1,2))
poverty <- poverty %>%
  mutate(treatment.factor = factor(if_else(treatment == "Before Payday", 1 , 0)))
poverty %>%
  ggplot(aes(x = stroop_time, fill = factor(treatment.factor))) +
  geom_histogram(position = 'identity', alpha = 0.5)
```

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




```
poverty %>%
  ggplot(aes(x = stroop_time, fill = factor(treatment.factor))) +
  geom_boxplot()
```




```
poverty$treatment <- as.factor(poverty$treatment)
```

Text Answer for 4b:

both the treatment(paid) and the no treatment(unpaid) groups have an overlap in their stroop times. their distribution is almsot entirely the same. similarly , with the side by side boxplots we can understand that the distribution of the variable strooptime is the same with the medians coinciding. The outliers are also approximately overlapping with each other.therefore, there isn't a significant difference in between both groups with their stroop times.

Question 5 (9 points)

5a (6 points)

Now, we will look at the relationship between general financial circumstances and cognitive performance. Produce two scatter plots, one for each of the two treatment conditions, showing the bivariate relationship between your *log-transformed* `cash` variable and `stroop_time` (place the `stroop_time` variable on the *vertical axis*). Be sure to title your graphs to differentiate between the `Before Payday` and `After Payday` conditions. Calculate the linear correlation between your *log-transformed* `cash` variable and `stroop_time` for each of the two treatment conditions. Do the associations between the log-transformed `cash` variable and cognitive performance appear to be linear? If yes, what is the direction and strength of this linear association? If no, what kind of association do you see between these two variables?

5b (3 points)

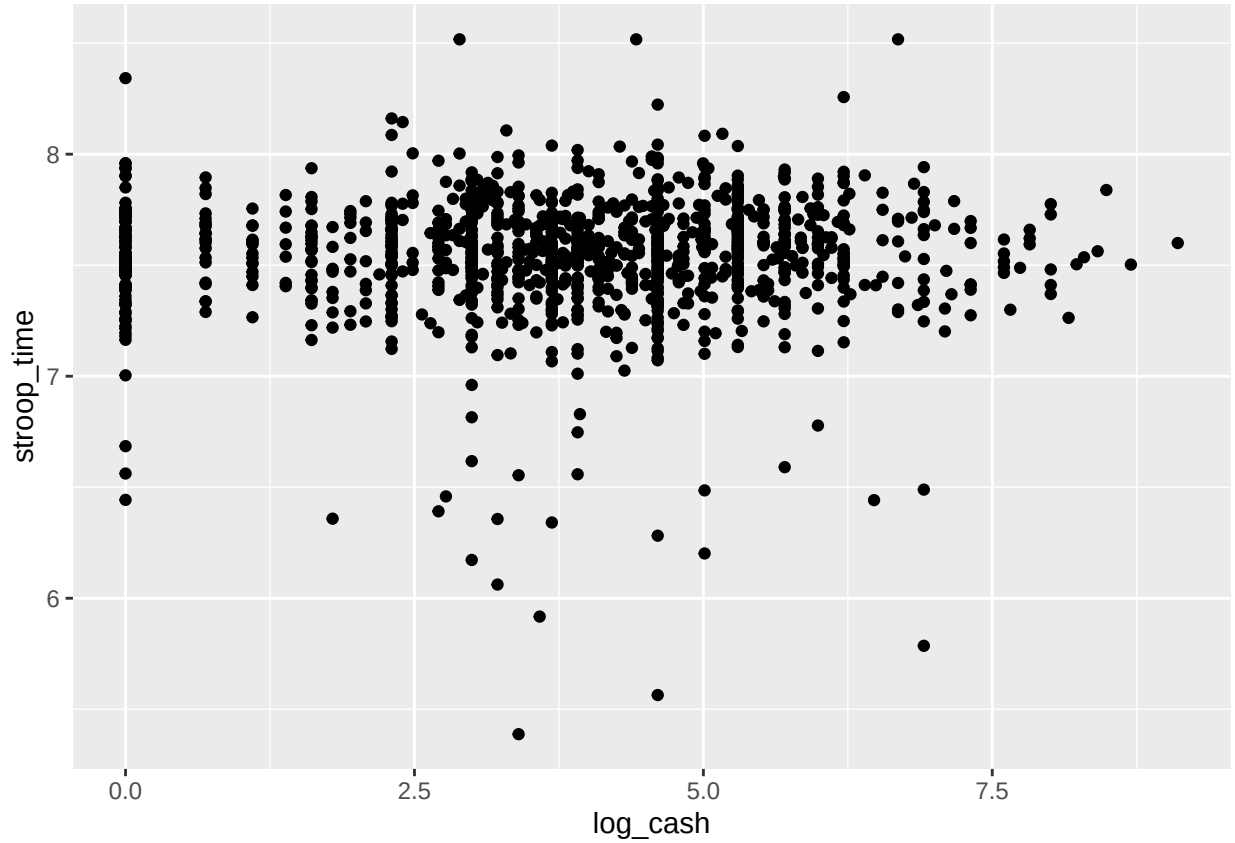
Using the scatterplots and the linear correlations as evidence, briefly comment on whether this evidence supports or contradicts the hypothesis that economic circumstances will influence cognitive performance.

Answer 5

5a

```
poverty_paid %>%
  ggplot(aes(x = log_cash, y = stroop_time)) +
  geom_point()
```

```
## Warning: Removed 114 rows containing missing values (`geom_point()`).
```

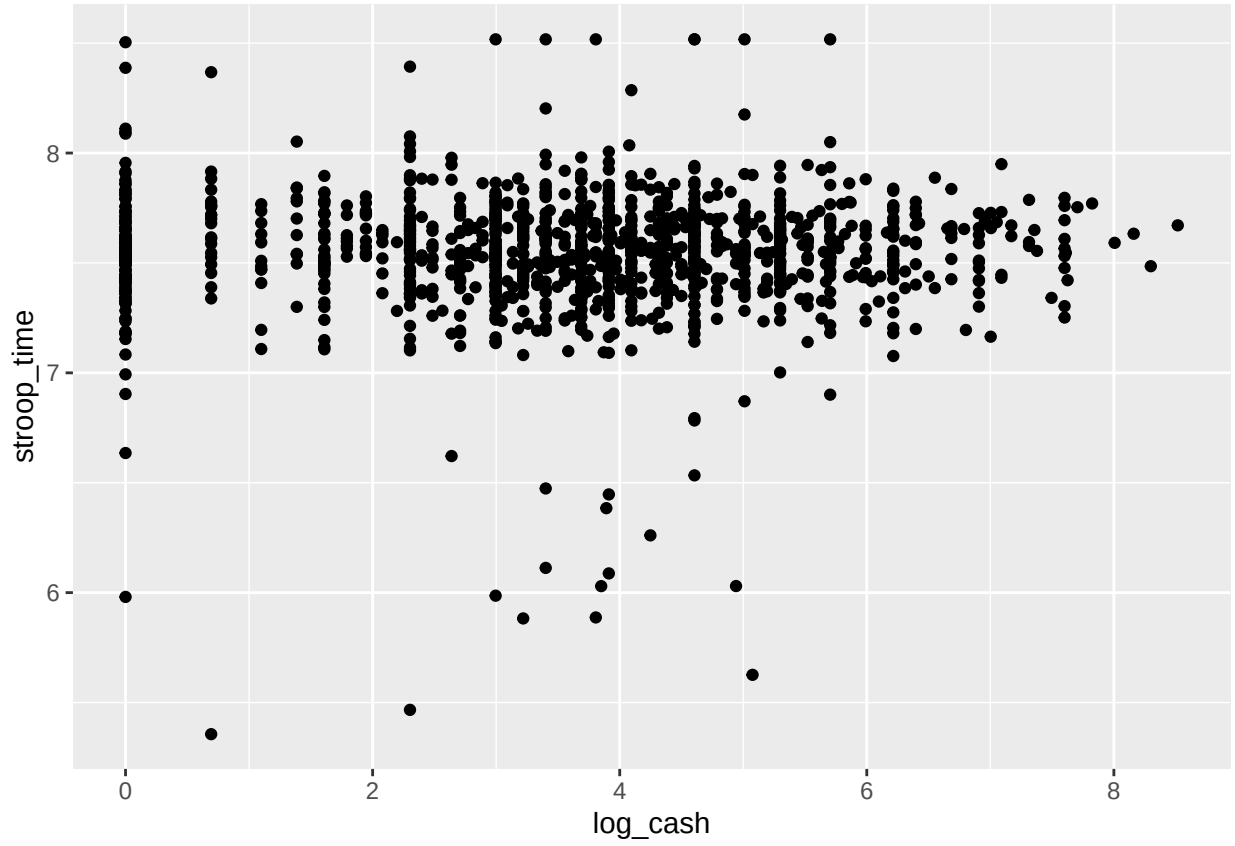



```
 labs( x = "LOG_CASH",
       y = " paid stroop time",
       labs = "paid"
  )
```

```
## $x
## [1] "LOG_CASH"
##
## $y
## [1] " paid stroop time"
##
## $labs
## [1] "paid"
##
## attr(,"class")
## [1] "labels"
```

```
poverty_unpaid %>%
  ggplot(aes(x = log_cash, y = stroop_time)) +
  geom_point()
```

```
## Warning: Removed 112 rows containing missing values (`geom_point()`).
```




```
labs( x = "LOG_CASH",
       y = " unpaid stroop time",
       labs = "unpaid"
  )
```

```
## $x
## [1] "LOG_CASH"
##
## $y
## [1] " unpaid stroop time"
##
## $labs
## [1] "unpaid"
##
## attr(,"class")
## [1] "labels"
```

```
poverty_paid_clean <- poverty_paid[!is.na(poverty_paid$log_cash),]
poverty_unpaid_clean <- poverty_unpaid[!is.na(poverty_unpaid$log_cash),]

cor(poverty_paid_clean$log_cash, poverty_paid_clean$stroop_time)
```

```
## [1] 0.02440794
```

```r
cor(poverty_unpaid_clean$log_cash, poverty_unpaid_clean$stroop_time)
```

```
## [1] 0.006404775
```

Text Answer for 5a:

the linear correlation between log transformed cash and the stroop time for before payday subjects : 0.0064 the linear correlation between log cash and the stroop time for after payday subjects is : 0.0244. The correlation for the after payday is slightly higher than the before payday subjects. the linear correlation indicates that the log_transformed cash variable and the stroop time for before payday subjects are not strongly related , but they have a positive correlation variable. the linear correlation indicates that the log_transformed cash variable and the stroop time for after payday are not strongly related but slightly more related than the before payday correlation. after payday: The linear correlation between the log transformed cash and the stroop time is near 0 and closer to zero. this could indicate that the two variables are in fact not related. there are outliers between log-cash value- 1.8 - 6.4 and have a lower stroop time between 1-7.Many values have a stroop time between 7-8 with the log_cash ranging from 0-8.25. Many subjects have zero as the log_cash variable but have a high stroop time variable of 7.4-8.2.

before payday: The linear correlation between the log transformed cash and the stroop time is near 0 and closer to zero. this could indicate that the two variables are in fact not related greatly. there are outliers between log-cash value- 2-6 and have a lower stroop time between 1-7.Many values have a stroop time between 7-8 with the log_cash ranging from 0-8.5. Many subjects have zero as the log_cash variable but have a high stroop time variable of 6.5-8.4. there are a few outliers with high stroop time values greater than 8.4.

Text Answer for 5b:

The researchers were interested in investigating whether or not being financially strained affects people's decision making and cognitive performance. We can infer from the graphs and the correlation coefficients that the 2 variables log_cash and stroop_time - that represents the cash at hand and their cognitive abilities respectively are not related and have a positive but close to zero correlation which would indicate that the evidence contradicts the hypothesis as it can be inferred that the financial situation of a subject would not greatly impact their cognitive decision making skills. ## Question 6 (16 points)

6a (7 points)

Create a z-score version of the `cash` variable. Use a summary command and a boxplot and then describe the distribution of the z-scores for this variable. What proportion of the observations have a z-score greater than 2? What proportion of the observations have a z-score greater than 3? How does this distribution differ from the distribution you described in question 2?

6b (9 points)

Create a z-score version of the log transformed `cash` variable. Use a summary command and a boxplot and then describe the distribution of the z-scores for this log transformed variable. What proportion of the observations have a z-score greater than 2 for this log transformed variable? What proportion of the observations have a z-score greater than 3 for this log transformed variable? How does this distribution differ from the distribution you described in question 3a?

Answer 6

6a

```
summary(poverty$cash)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    15.0    49.5   169.0   136.2  9000.0     226
```

```
sd(poverty_cash$cash)
```

```
## [1] 451.283
```

```
poverty_cash['cash.z.score'] <- (poverty_cash$cash - mean(poverty_cash$cash))/sd(poverty_cash$cash)
mean(poverty_cash$cash.z.score)
```

```
## [1] 2.84412e-17
```

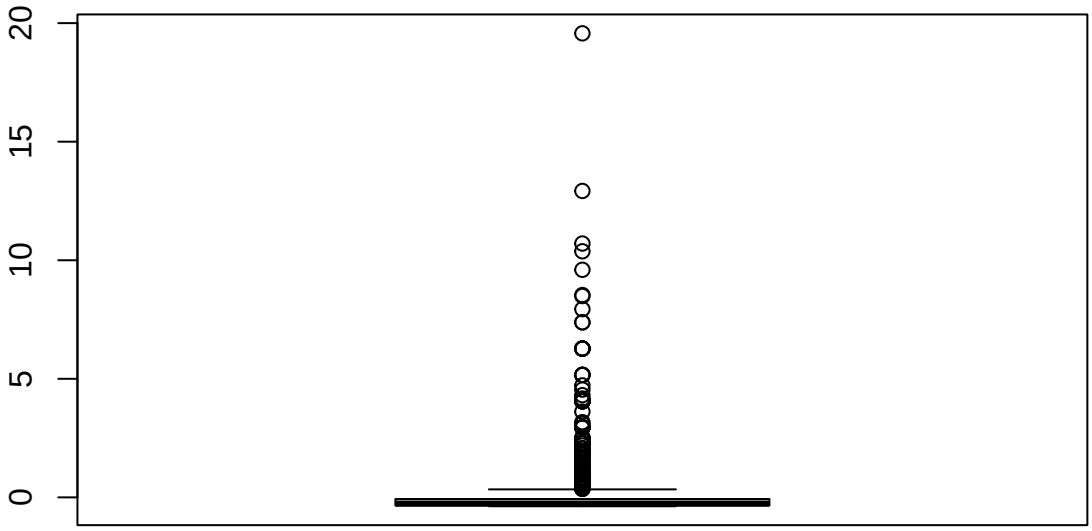
```
sd(poverty_cash$cash.z.score)
```

```
## [1] 1
```

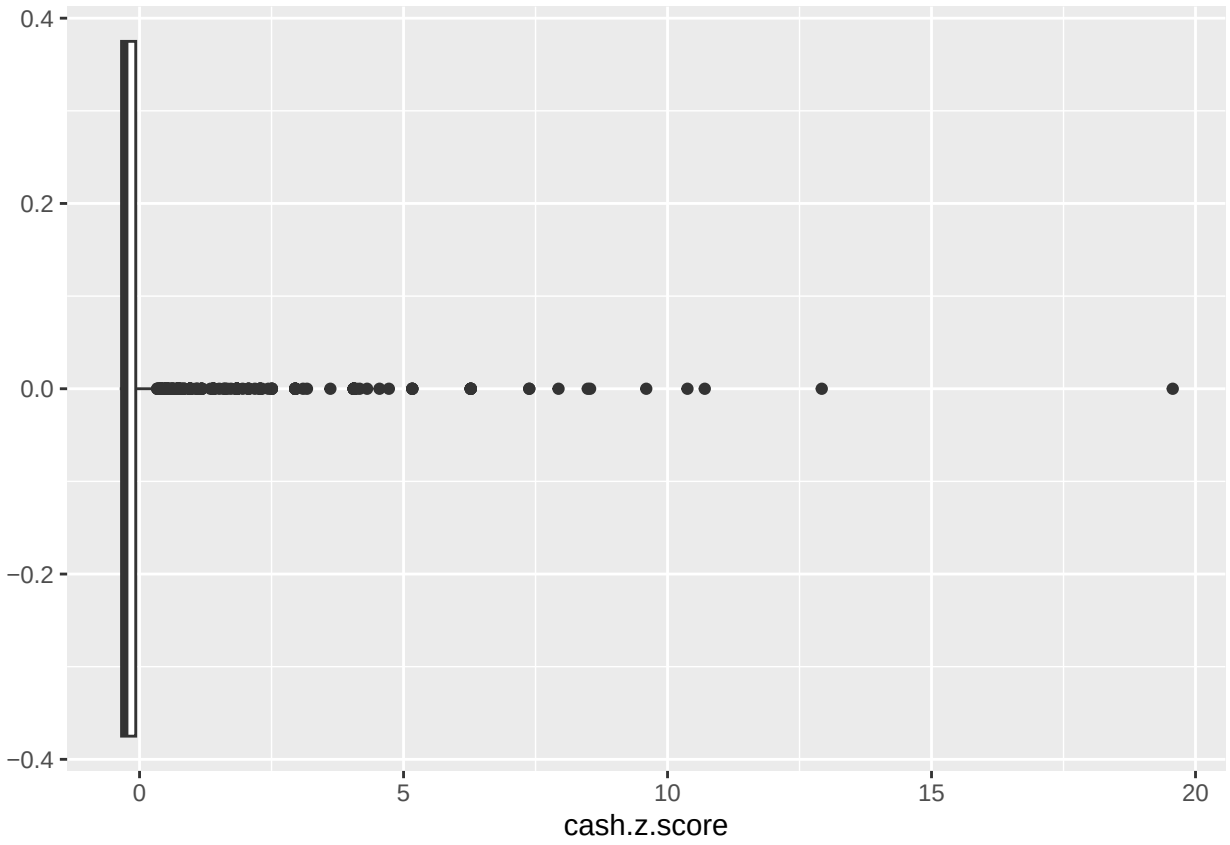
```
summary(poverty_cash$cash.z.score)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.37450 -0.34126 -0.26481  0.00000 -0.07258 19.56865
```

```
boxplot(poverty_cash$cash.z.score)
```




```
poverty_cash %>%
  ggplot(aes(x = cash.z.score)) +
  geom_boxplot()
```




```
mean(poverty_cash$cash.z.score > 2)
```

```
## [1] 0.02823241
```

```
mean(poverty_cash$cash.z.score > 3)
```

```
## [1] 0.01718494
```

```
mean(poverty_cash$cash > 2)
```

```
## [1] 0.891162
```

```
mean(poverty_cash$cash > 3)
```

```
## [1] 0.8801146
```

Text Answer for 6a:

the range of this z score variable for cash is: -0.37- 19.56. the sd = 1. the mean is 0. the median is -0.26 and 50 percent of the values lie between the 1st and the 3rd quartiles: -0.34 and -0.072. the cash z score values have a scale from 0 -20 on the boxplot and the one for cash variable ranges from 0-7500. the ditribution of outliers is similar to what we observe in the cash variable in Q2 as compared to the z scored tranformed variable of cash . they both have one outlier several IQRS away from the tail.

- 2.8 % of the z score cash variable values are more than 2 standard deviations above the mean and 1.7 % are above 3 standard deviations above the mean. with the cash variable we can observe that 89% of the z score values are more than 2 standard deviations above the mean. We can observe that 88 % of the z score cash values are 3 sd's above the mean. the two distributions are similar. the outliers are attached to the tail end.which makes it right skewed ### 6b

```
summary(poverty$log_cash)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.708   3.902   3.684   4.914   9.105     226
```

```
sd(poverty_cash$log_cash)
```

```
## [1] 1.842889
```

```
poverty_cash['logcash.z.score'] <- (poverty_cash$log_cash - mean(poverty_cash$log_cash))/sd(poverty_cash

mean(poverty_cash$logcash.z.score)
```

```
## [1] -1.001182e-16
```

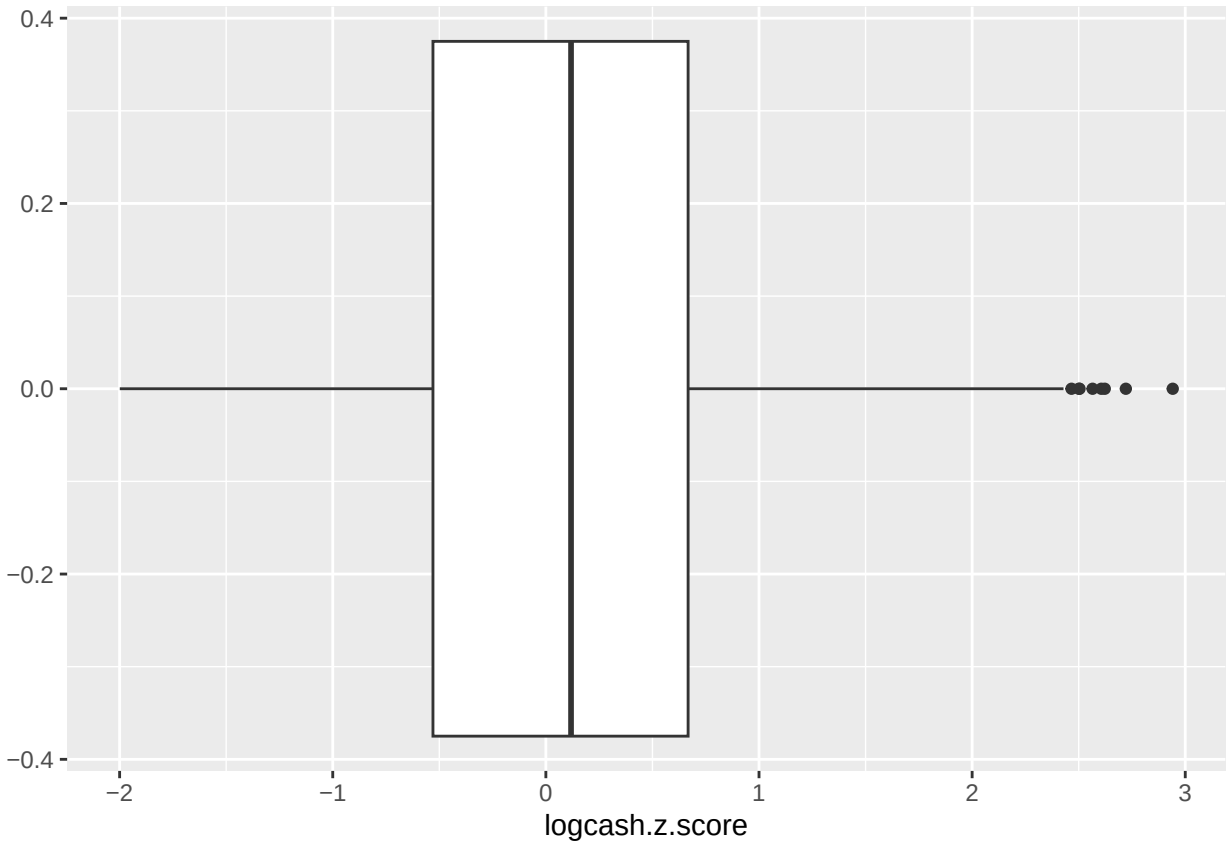
```
sd(poverty_cash$logcash.z.score)
```

```
## [1] 1
```

```
summary(poverty_cash$logcash.z.score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.9993 -0.5298  0.1180  0.0000  0.6674  2.9413
```

```
poverty_cash %>%
  ggplot(aes(x = logcash.z.score)) +
  geom_boxplot()
```




```
mean(poverty_cash$logcash.z.score > 2)
```

```
## [1] 0.01677578
```

```
mean(poverty_cash$logcash.z.score > 3)
```

```
## [1] 0
```

```
mean(poverty_cash$log_cash>2)
```

```
## [1] 0.8306056
```

```
mean(poverty_cash$log_cash>3)
```

```
## [1] 0.6661211
```

Text Answer for 6b:

the range of this z score variable for log_cash is: score scale is from -1.99- 2.9. the sd = 1. the mean is 0. the median is 0.11 and 50 percent of the values lie between the 1st and the 3rd quartiles: -0.52 and 0.66. The log_cash z score values have a scale from -2 to 3 on the boxplot and the one for log_cash variable is 0-7.5 on the X axis. The distribution of outliers the distribution is identical both boxplots are symmetrical

and have outliers attached to the tail end of the distribution.They both have one outlier several 1/4 th an IQR away from the tail.

1.6 % of the z score log cash variable values are more than 2 standard deviations above the mean and 0 % are more than 3 standard deviations above the mean. we can observe that the log cash variable from part 3a has 83% more than 2 standard deviations above the mean. We can observe that 66 % of log cash values in 3a are 3 sd's above the mean. the two distributions are similar. the boxplot is symmetrical with outliers on the whiskers and they are between 1- 1.5 IQR's away.