# 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. Under this design, the researchers believed that the respondents of the `Before Payday` group would be more likely to be financially strained than those of the `After Payday` group. 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 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 |
| `accts_amt` | Amount in checking and saving accounts |
| `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 (5 points)

### 1a

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

### 1b

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

### 1c

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

## Answer 1

1a. What is the impact of having a financial constraint in the form of not receiving a paycheck relative to the financial stability of the individuals who did get a paycheck on the individuals cognitive ability (stroop time) of the individuals who did get a pay check?

1b. What the participants cognitive ability if they have not yet received a pay check and what is the participants cognitive ability if they have received a pay check?

1c. The strategy used to estimate the MCF is pre-post. We are looking at the results before payday and we are looking at the results after the payday. The assumption that is required for the estimate to be unbiased, is that the stroop-time would stay the same if the individual did not receive a paycheck.

## Question 2 (10 points)

### 2a

Load the `poverty.csv` data set. Use histograms and boxplots to examine the univariate distributions of the two financial resources measures: `cash` and `accts_amt`. Calculate the mean and standard deviation, median, quartiles, and IQR for the two financial resources measures.

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

### 2b

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 `financial resource`cash' variable. Create a boxplot and histogram for the transformed variable.

Describe the log transformed variable's distributions based on the figures and the summary statistics.

### 2c

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.*

```
# You will need ggplot2 to make plots.
require("ggplot2")
```

```
## Loading required package: ggplot2
```

```
require("tidyverse")
```

```
## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Feel free to load any other packages you want

# Load the poverty data; remember to put the csv in a data sub-folder!
poverty <- read.csv("data/poverty.csv") %>%
  na.omit()# remove rows with missing data
```

```
### Log transformation code for 2b
# Feel free to move this code where you need it.

# Log-transform money in cash
poverty$log_cash <- poverty$cash
```

```r
poverty$log_cash[poverty$log_cash == 0] <- 1
poverty$log_cash <- log(poverty$log_cash)
```

## Answer 2

Your answers for Question 2 go here! 2a. Calculate the mean and standard deviation, median, quartiles, and IQR for the two financial resources measures. Describe the variables' distributions based on the figures and the summary statistics.

Cash Var:

The min value is 0. The Q1 value is 17, The Q3 value is 148. The IQR = 148-14= 131. The standared deviation is 464.0278. The max value is 9000. The histogram for the cash variable is unimodal and right skewed. We have a spike at 0. The range of our data is between 0 and 9000.

accts_amt:

The min value is 0. The Q1 value is 176, The Q3 value is 5000. The IQR = 5000-176=4824 . The standard deviation is 13517.69. The max value is 95000. The histogram for the accts_amt variable is unimodal and right skewed. We have a spike at 0. The range of our data is between 0 and 95000.

2b. Describe the log transformed variable's distributions based on the figures and the summary statistics.

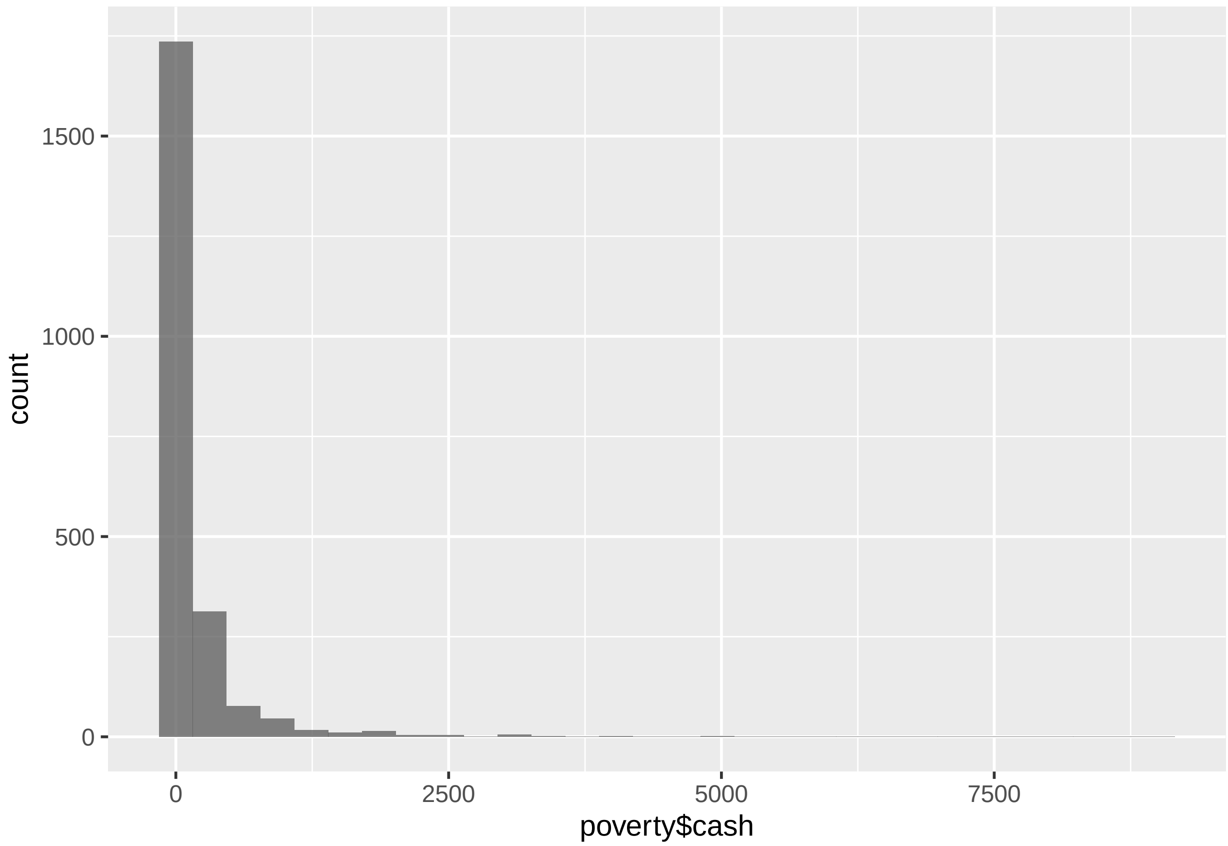
Cash log Var:

The min value is 0. The Q1 value is 2.833, The Q3 value is 4.997. The IQR = 4.997- 2.833= 2.164. The standard deviation is 1.814489. The max value is 9.105. The histogram for the cash log variable is unimodal and symmetrical. The range of our data is between 0 and 9.105.

2c. Advantages to transforming the data is that we can compare the data at a more uniform level because it is shifted more towards the mean and the disadvantage to transforming is that the data is not showing the chart based on the original representation of the data.

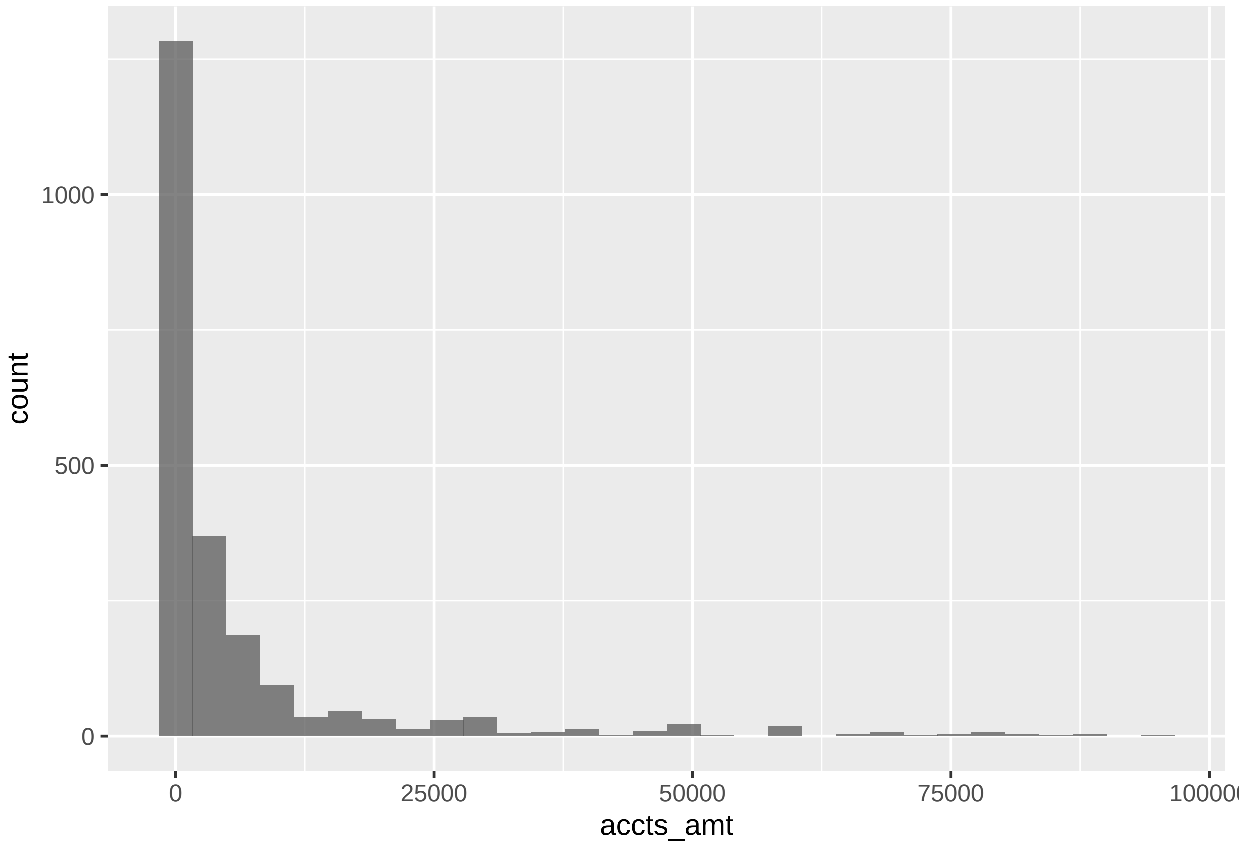
```r
# Your code goes here!



poverty %>%
  ggplot(aes(x = poverty$cash)) +
  geom_histogram( alpha = 0.75)
```

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

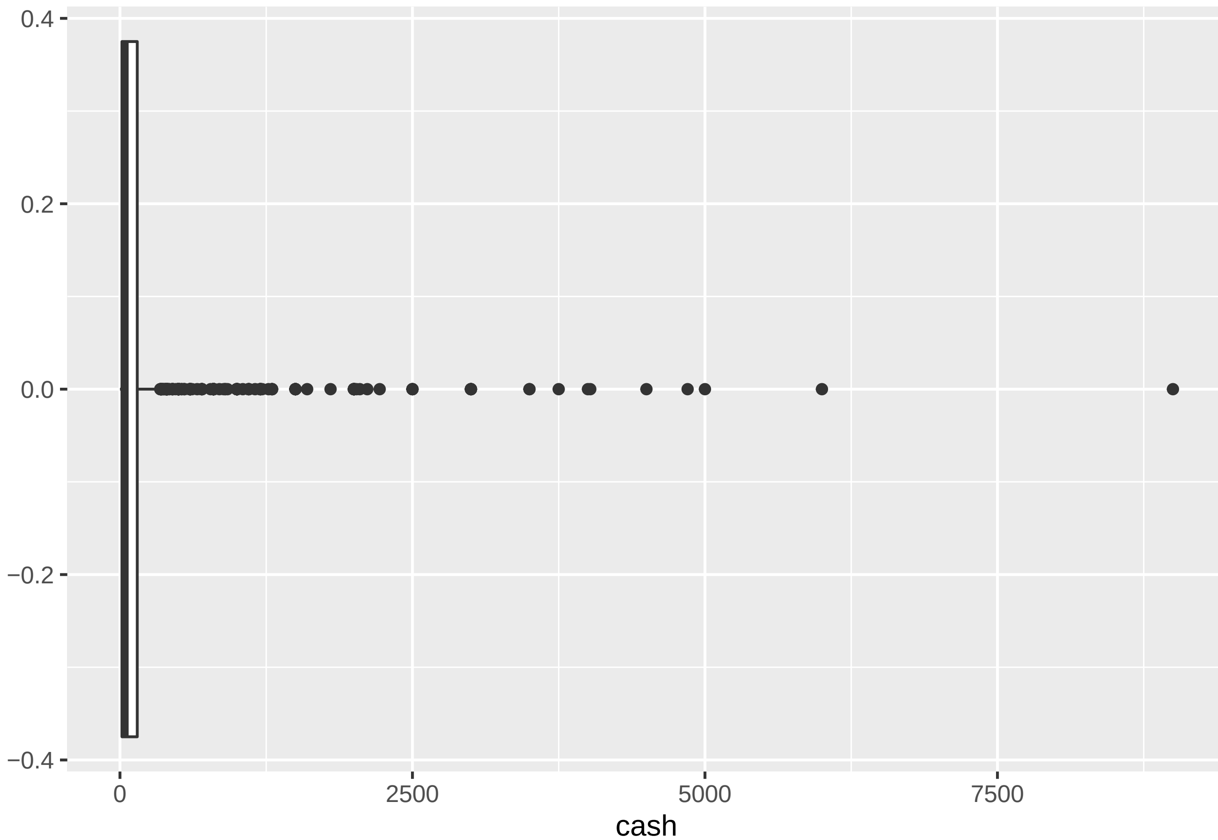



```
poverty %>%
  ggplot(aes(x = accts_amt)) +
  geom_histogram( alpha = 0.75)
```

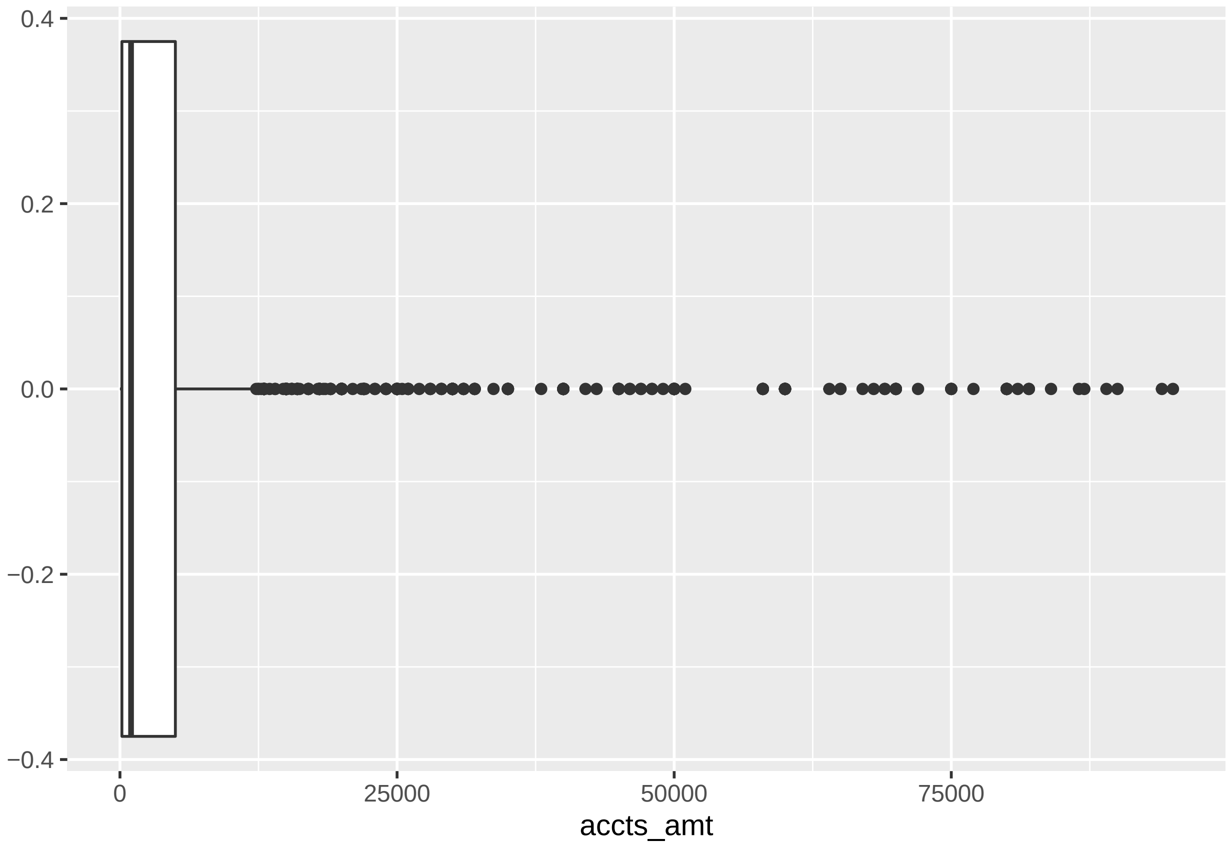
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
poverty %>%
  ggplot(aes(x = cash))+
  geom_boxplot()
```




```r
poverty %>%
  ggplot(aes(x = accts_amt))+
  geom_boxplot()
```




```
sd(poverty$cash)
```

```
## [1] 464.0278
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    17.0    50.0   174.3   148.0  9000.0
```

```
sd(poverty$accts_amt)
```

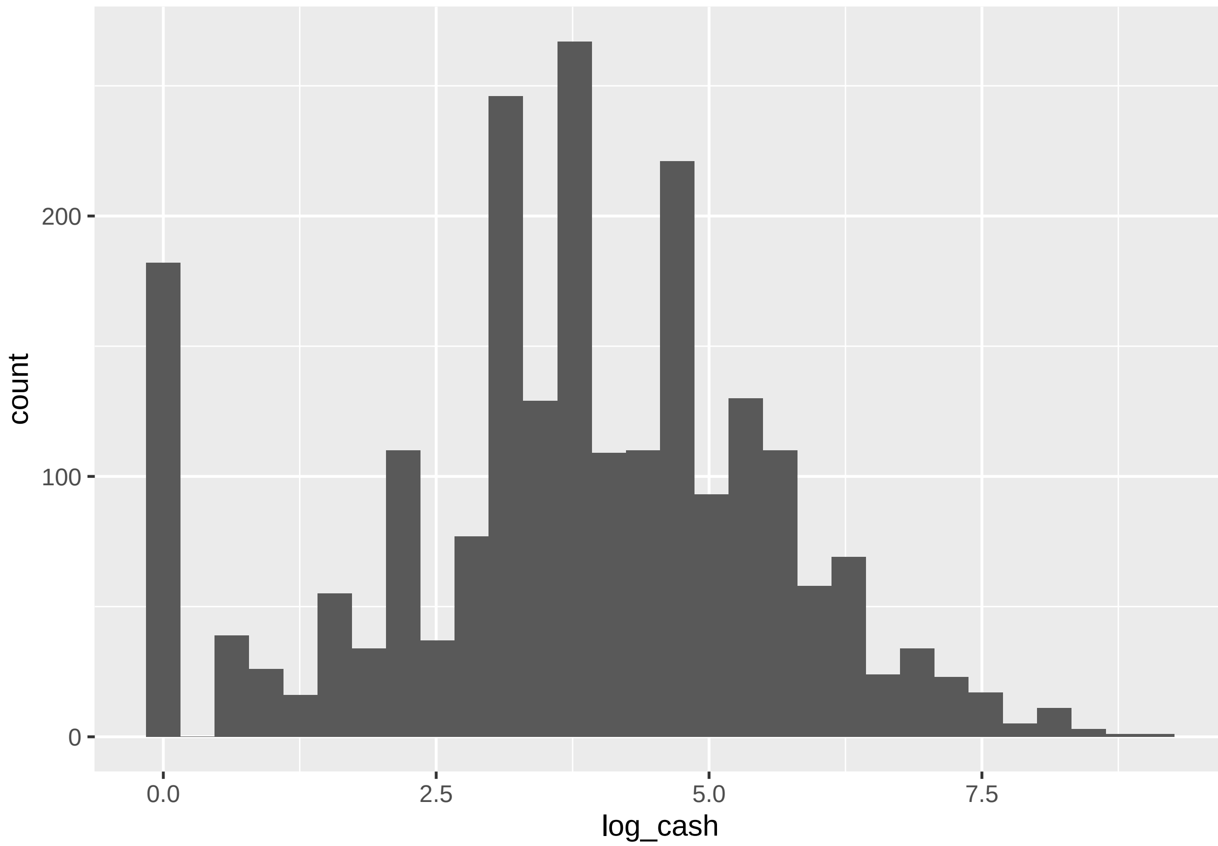
```
## [1] 13517.69
```

```
summary(poverty$accts_amt)
```

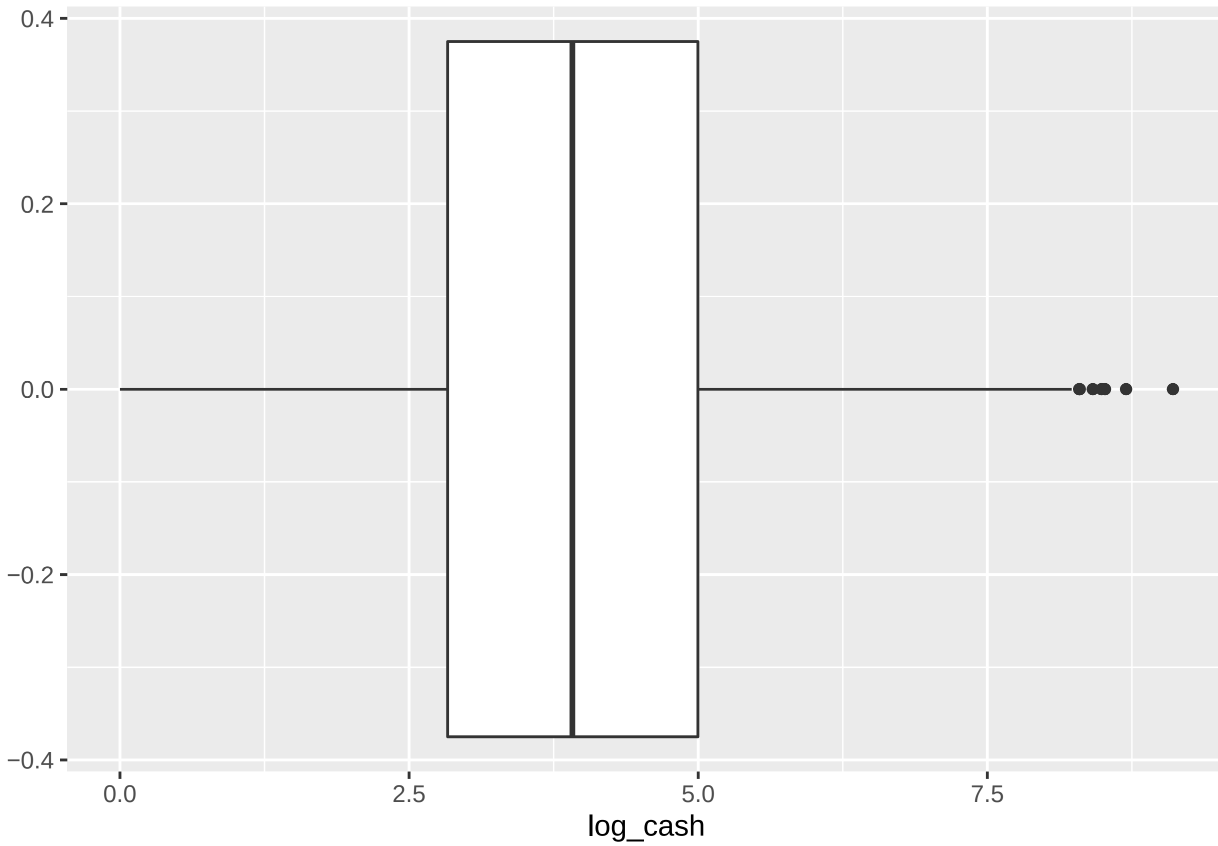
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     176    1000    6211    5000   95000
```

```
poverty$log_cash <- poverty$cash
poverty$log_cash[poverty$log_cash == 0] <- 1
poverty$log_cash <- log(poverty$log_cash)

poverty  %>%
  ggplot(aes(x = log_cash)) +
  geom_histogram( )
```

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




```r
poverty  %>%
  ggplot(aes(x = log_cash))+
  geom_boxplot()
```




```r
sd(poverty$log_cash)
```

```
## [1] 1.814489
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.833   3.912   3.742   4.997   9.105
```

## Question 3 (10 points)

### 3a

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?

### 3b

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?

## Answer 3

Your answers for Question 3 go here!

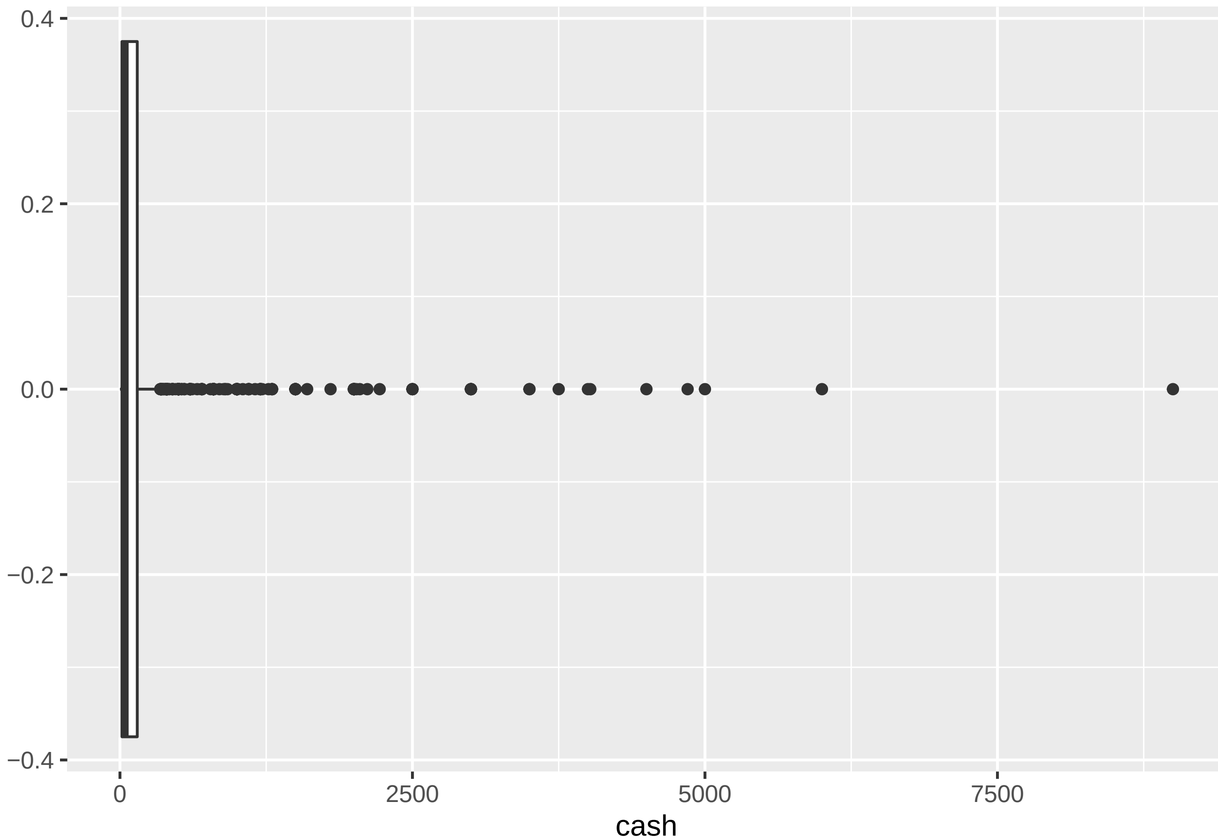
3a. The proportions of z-scores greater than 2 is 0.02771569. The proportions of z-scores greater than 3 is 0.01743406.

3b. The proportions of z-scores greater than 2 is 0.01743406. The proportions of z-scores greater than 3 is 0.

```r
z.score = ((poverty$cash - mean(poverty$cash))/sd(poverty$cash))
summary(z.score)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.37569 -0.33905 -0.26793  0.00000 -0.05674 19.01971
```

```r
poverty %>%
  ggplot(aes(x = cash))+
  geom_boxplot()
```




```r
mean(if_else(z.score > 2, 1, 0))
```

```
## [1] 0.02771569
```

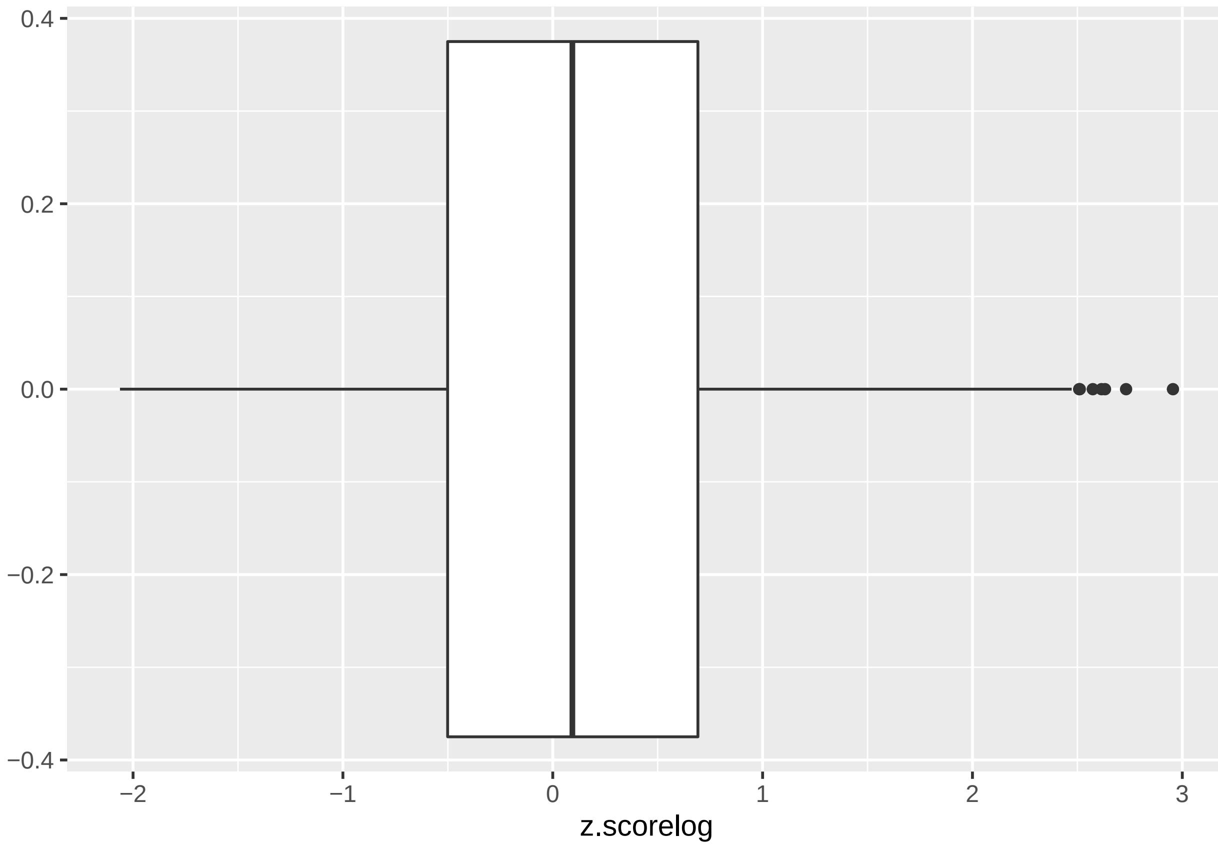
```r
mean(if_else(z.score > 3, 1, 0))
```

```
## [1] 0.01743406
```

```r
#.........................................................

z.scorelog = ((poverty$log_cash - mean(poverty$log_cash))/sd(poverty$log_cash))
summary(z.scorelog)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.06241 -0.50097  0.09358  0.00000  0.69165  2.95552
```

```r
poverty %>%
  ggplot(aes(x = z.scorelog))+
  geom_boxplot()
```




```r
mean(if_else(z.scorelog > 2, 1, 0))
```

```
## [1] 0.01743406
```

```r
mean(if_else(z.scorelog > 3, 1, 0))
```

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

## Question 4 (5 points)

### 4a

To what extent was there balance between the `Before Payday` group and the `After Payday` group on having a low annual income? Use appropriate summary statistics to determine this.

### 4b

What are the implications of this for the internal validity of the study?

## Answer 4

4a. The data for before the pay and after the payday was very closely related.

4b. The average of both before and after treatment groups have a very similar distribution for the low annual income participants. The data that we have gotten to represent the low income participants is shown with the group 1. In the before group the mean is .40 and the after group is around .38 and with this we can see that both groups are relatively the same. If we view the study using the pre-post structure we can say that the estimate of the missing counter-factual is correct because the averages are similar. Due to that we can say that the internal validity for the study is strong.

```r
## Look at balance of low annual income

prop.table(table(poverty$income_less20k, poverty$treatment), margin=2)
```

```
##
##     After Payday Before Payday
##   0    0.6203111     0.5961538
##   1    0.3796889     0.4038462
```

```r
#1 is the low income group 4
```

## Question 5 (6 points)

### 5a

Now, let's focus on the primary outcome of interest for this study - cognitive performance. Let's study 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. Also, calculate the standard deviation and IQR of the outcome for each treatment group. What do these summary statistics tell you about differences in cognitive performance across the two treatment conditions?

### 5b

Now compare the time it took subjects to complete the cognitive test across the `Before Payday` and `After Payday` groups using overlaid histograms. Based on your plot, do you think there is a meaningful difference in the time it took participants from each group to complete the test?

## Answer 5

5a.

Before Payday:

Mean:7.541 Med:7.555 IQR: 0.2498953 SD: 0.2700677

After Payday: Mean:7.545 Med:7.57 IQR: 0.2509988 SD: 0.2705974

The summery statistic tells us that the cognitive ability of the participants in the study group before payday and after payday is very similar. The mean, medium and range for the data is also very close to each other. For both the before and after groups the standard deviation is very small so our data isnt deviating that far from the mean. The IQR for our data sets for both the groups are in the 50% range is around .27. With all this information we can conclude that the groups are not affecting the cognitive ability of the participants in the group.

5b.

We can see from the histogram overlap that we don't have a meaningful difference for the peoples cognitive ability from their financial status. The data is very closely related. The histogram representing the cognitive ability of the participants is symmetric and unimodal for both the before and after payday groups.

```r
## Calculations for the questions on summary statistics

summary(poverty$stroop_time[poverty$treatment == "Before Payday"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.356   7.431   7.555   7.541   7.680   8.517
```

```r
IQR(poverty$stroop_time[poverty$treatment == "Before Payday"])
```

```
## [1] 0.2498953
```

```r
sd(poverty$stroop_time[poverty$treatment == "Before Payday"])
```

```
## [1] 0.2700677
```

```r
summary(poverty$stroop_time[poverty$treatment == "After Payday"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.388   7.440   7.570   7.545   7.691   8.517
```

```r
IQR(poverty$stroop_time[poverty$treatment == "After Payday"])
```

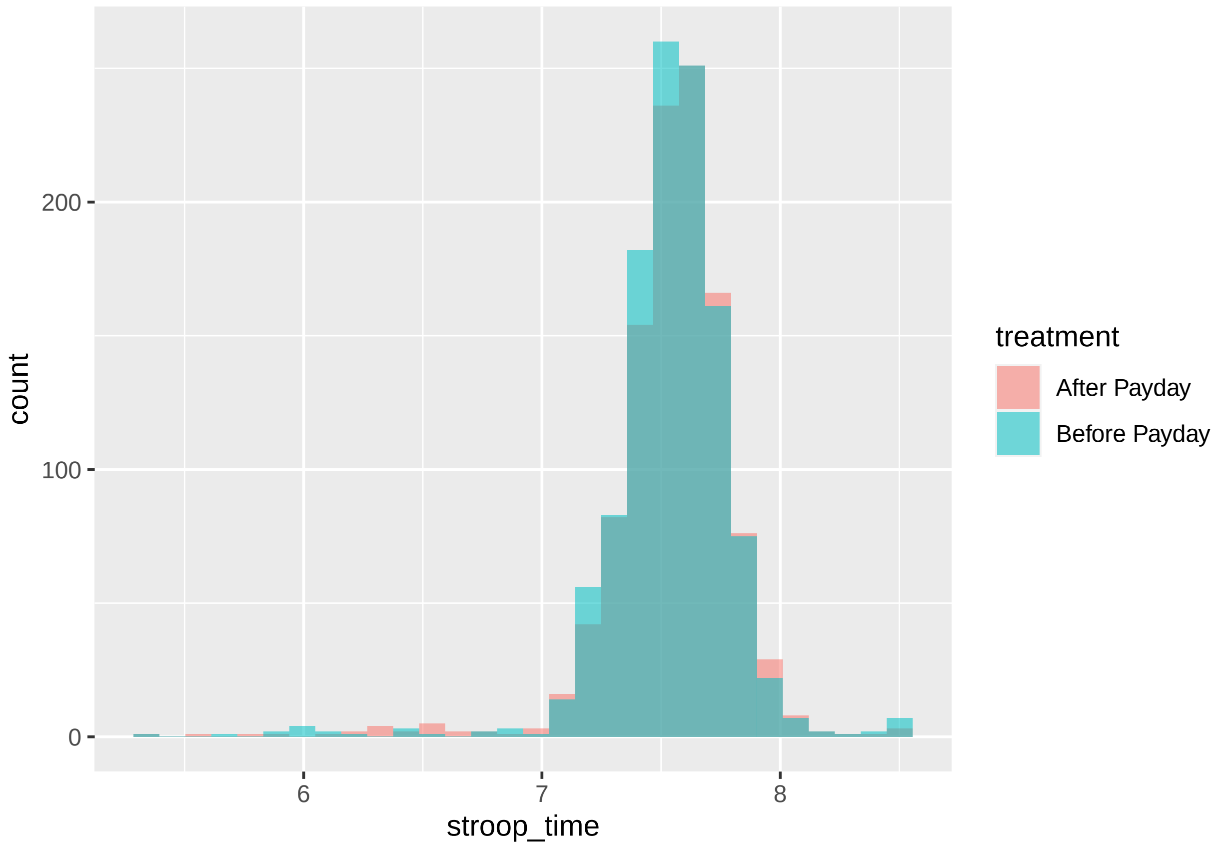
```
## [1] 0.2509988
```

```r
sd(poverty$stroop_time[poverty$treatment == "After Payday"])
```

```
## [1] 0.2705974
```

```r
## Histograms!

poverty %>%
  ggplot(aes(x = stroop_time, fill = treatment)) +
  geom_histogram(position = 'identity', alpha = 0.55)
```

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



## Question 6 (8 points)

### 6a

Now, we will look at the relationship between general financial circumstances and cognitive performance. Produce two scatter plots side by side, 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 y-axis [y-axis is 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 experimental 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?

### 6b

Using the scatterplots and the linear correlations as evidence, briefly comment on your results in light of the hypothesis that changes in economic circumstances will influence cognitive performance.

BONUS: Produce one scatter plot showing the bivariate relationship between your *log-transformed* `cash` variable and `stroop_time` for **each experimental condition**, using different colors (or different symbol shapes) to identify each experimental condition. Note - before we created two separate scatterplots for this information; now, we want you to make just one, but use colors or symbols to differentiate between observations in the `Before Payday` and `After Payday` treatment arms.

6a. We dont have a weak (Negative) or strong (Positive) correlation because we see no linear trend in the data. We see clustered data points but no association between the variables because we cannot predict where the y and x value will fall under. The stroop time (cognitive ability) and the log cash is close to 0 also demonstrating to us that we have little to no association between the variables.

6b.

Based on the results we can see and that we dont have a strong or a weak correlation on the data, we can say that that the economic circumstance will have little or no influence on cognitive performance.

## Answer 6

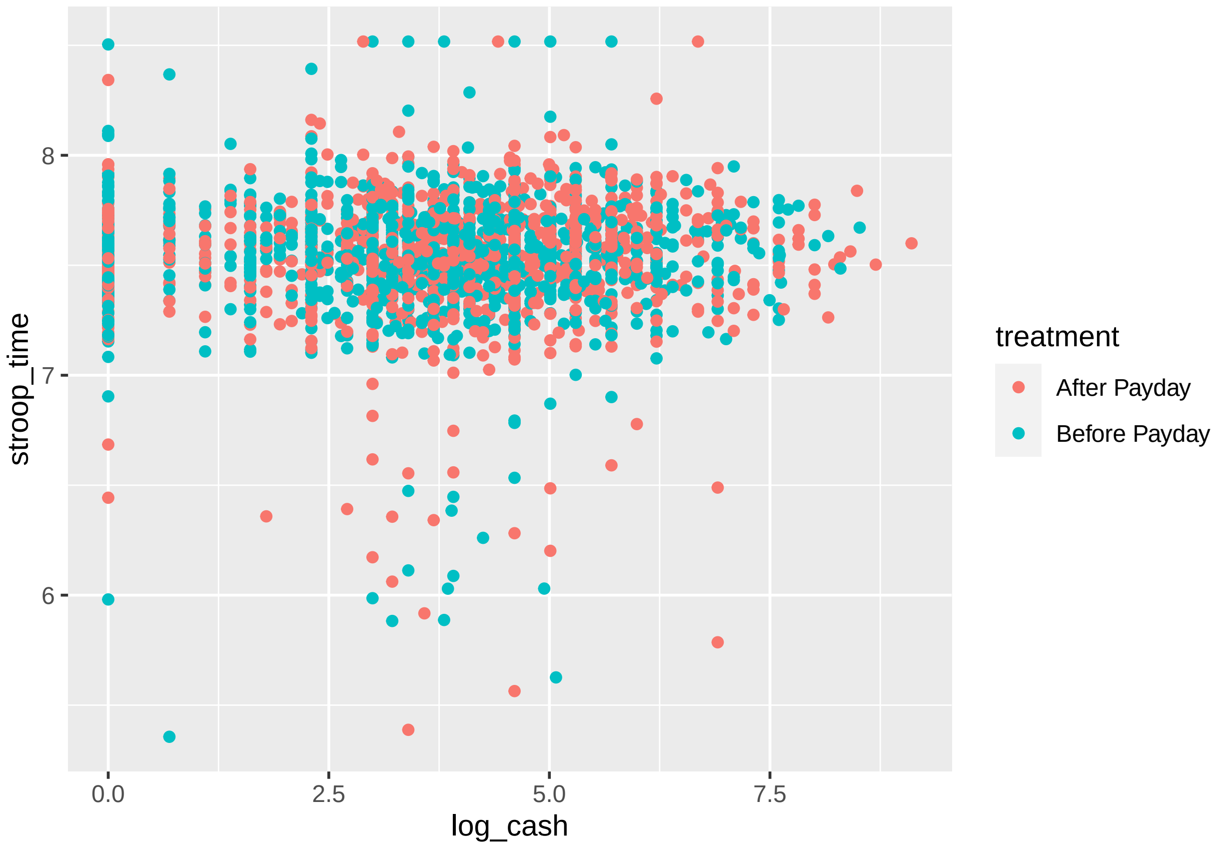
```r
## Calculations
cor(poverty$log_cash[poverty$treatment == "Before Payday"], poverty$stroop_time[poverty$treatment == "Be
```

```
## [1] -0.006017568
```

```r
cor(poverty$log_cash[poverty$treatment == "After Payday"], poverty$stroop_time[poverty$treatment == "Af
```

```
## [1] 0.02100648
```

```r
## Graphs
poverty %>%
  ggplot(aes(x=log_cash , y=stroop_time, color = treatment)) +
  geom_point()
```




```
#facet_wrap(~treatment)
```