S58_data_test

S58

05/02/2022

```
library(tidyverse)
library(haven)
library(moments)
library(expss)
library(stargazer)
library(ggpubr)
```

1. DATA PREPARARTION

a.) Load 'endline' data and other data sets.

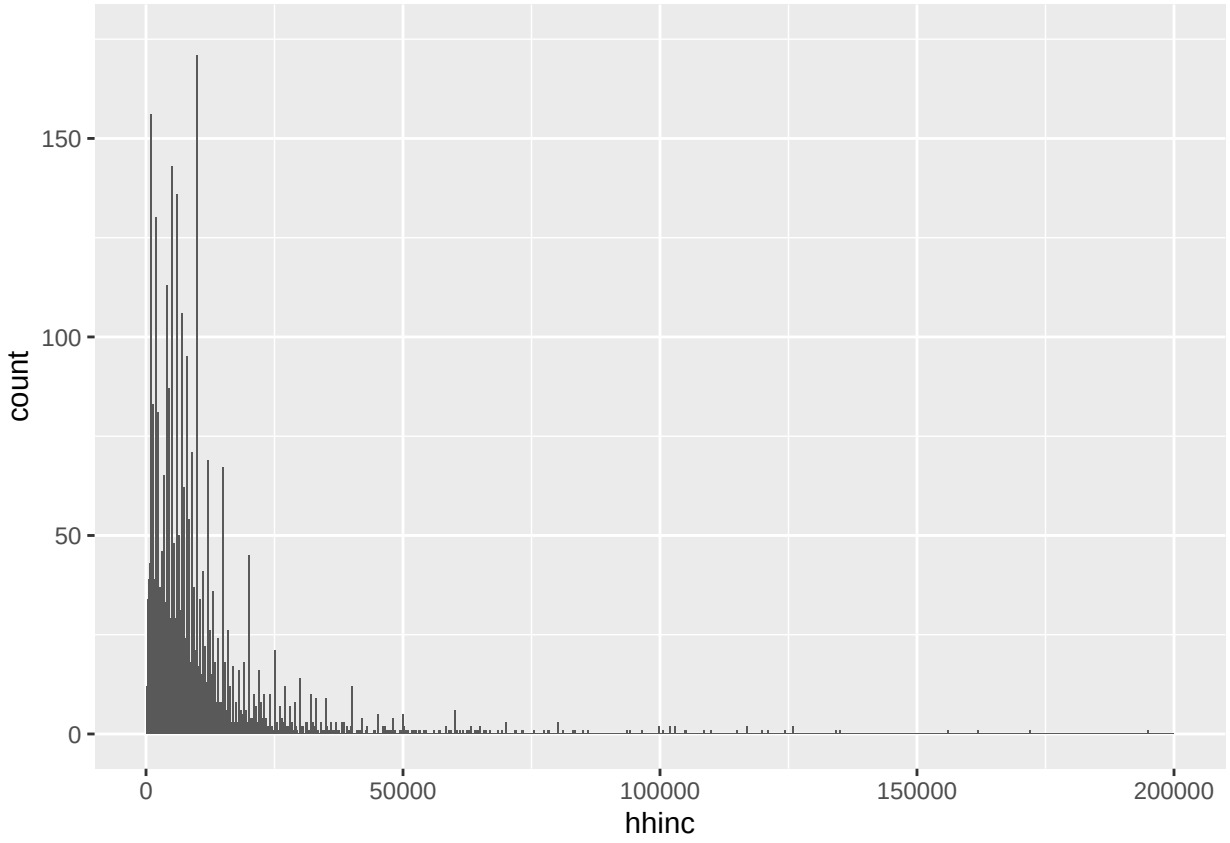
```
baseline_controls <- read_dta("Stata_Test_2020/baseline_controls.dta")
endline <- read_dta("Stata_Test_2020/endline.dta")
treatment_status <- read.csv("Stata_Test_2020/treatment_status.csv")
```

b.) Recode 'household' debt and 'income' variable as numerics and replacing "None" with '0'.

```
# replacing '.' with NA and None' with 0.
## for varibale 'totformalborrow_24'

endline$totformalborrow_24 <-
  str_replace_all(endline$totformalborrow_24, c("\\." = "NA", "None" = "0")) %>%
  as.numeric(endline$totformalborrow_24)

## for varibale ''
endline$totinformalborrow_24 <-
  str_replace_all(endline$totinformalborrow_24, c("\\." = "NA", "None" = "0")) %>%
  as.numeric(endline$totinformalborrow_24)

## for varibale 'hhinc'
endline$hhinc <-
  str_replace_all(endline$hhinc, c("\\." = "NA", "None" = "0")) %>%
  as.numeric(endline$hhinc)
```

c.) financial status of households[1]

All the variables are highly (positively) skewed. The skewness of 'household income' (hhinc) is 51.009905, 'total formal borrowing income' (totformalborrow_24) is 9.7285941, and for 'total informal borrowing' (totinformalborrow_24) it is 4.8908976. This hints at the high inequality among the HHs. The average value for the households income is 1.1809389×10^4, Whereas, the median value is 6000, showing the skewness of the income distribution among the households. The attached graph shows the skewness of the income.



Additionally, average formal borrowing is 6.4382051×10^4 which is greater than average informal borrowing at 4.0921066×10^4. It is might be due to the intervention that the formal borrowing shows a higher value than that of the informal one.

d.) top code household 'debt' and 'income' variables.

```
# household income
top_val_hhinc <-3*sd(endline$hhinc, na.rm = T)+mean(endline$hhinc, na.rm = T)
endline$hhinc[endline$hhinc > top_val_hhinc] <- top_val_hhinc

# household total formal borrowing
top_val_fb <-3*sd(endline$totformalborrow_24, na.rm = T)+
  mean(endline$totformalborrow_24, na.rm = T)
endline$totformalborrow_24[endline$totformalborrow_24 > top_val_fb] <- top_val_fb
```

[1] All the results are dynamically generated. The codes are available in the rmd file but has been supressed from final report.

```
# household total informal borrowing
top_val_ifb <-3*sd(endline$totinformalborrow_24, na.rm = T)+
  mean(endline$totinformalborrow_24, na.rm = T)
endline$totinformalborrow_24[endline$totinformalborrow_24 > top_val_ifb] <- top_val_ifb
```

e.) labeling variables

```
endline <- apply_labels(endline,
                        hhinc = "household income",
                        totformalborrow_24 = "total formal borrowing",
                        totinformalborrow_24 = "total informal borrowing")
```

f.) top-coding values

The reason to top-code the data is to treat the outliers. The presence of outliers skews the distribution and makes the inferences unreliable. In presence of outliers, another kind of treatment could be to implement a logarithmic transformation of the variable. This is bound to reduce the influence of outliers on the analysis. Yet another check could be to separately study the outlier variable in their context. The knowledge of other attributes associated with these outlier observations could inform us better to take further steps. Additionally, If this is a product of sampling error then deletion could be adopted.

g.) total borrowed amount

```
endline <- endline %>% mutate(tot_borrw = totformalborrow_24+totinformalborrow_24)

endline <- apply_labels(endline, tot_borrw = "total borrowing")
```

h.) merging 'endline' dataset with 'treatment_status' data set.

```
endline_treat <- endline %>% left_join(treatment_status, by = "group_id")
```

i.) poverty line dummy

```
## creating the variable 'daily_hhinc'.
endline_treat <- endline_treat %>% mutate(daily_hhinc = hhinc/30)

endline_treat <- apply_labels(endline_treat, daily_hhinc = "daily household income")

## creating the variable "daily household income per capita" as 'daily_hhinc_pc'.

endline_treat <- endline_treat %>% mutate(daily_hhinc_pc = daily_hhinc/hhnomembers)

endline_treat <- apply_labels(endline_treat, daily_hhinc_pc = "daily per capita income")
```

```
## creating dummy variable 'poverty_dummy' to indicate whether the household is below poverty line.
## Households with per day per capita income less than 26.995 have been assigned
## a 'poverty_dummy' value 1.

endline_treat <- endline_treat %>%
  mutate(poverty_dummy = ifelse(endline_treat$daily_hhinc_pc >= 26.995, 0, 1))

endline_treat <- apply_labels(endline_treat, poverty_dummy = "1 indicates below poverty line")

## Yes. There are missing values in the variable 'poverty_dummy'

miss_pov_dum <- sum(is.na(endline_treat$poverty_dummy))
```

There are 4 missing values in 'poverty_dummy'

j.) strengths and limitations of 'poverty_dummy'

The poverty line is an absolute measure of poverty and makes a comparison between different groups (even countries) easier. However, this does not take into consideration the differences in the cost of living for different groups (or countries). Also, households just above and just below the poverty line do not differ much in their per day per capita earning but are assigned different poverty indicators. Additional information on household consumption, access to financial services, schooling and health services, etc could help create a better classification of poors. Questions that can help collect this information could enrich the survey.

k.) merging 'endline_treat' dataset with 'baseline_controls' data set.

```
## This indicates if the hh was in endline survey.
endline_treat$end <- 2

## This indicates if the hh was in baseline survey.
baseline_controls$base <- 1

## Joining 'endline' and 'baseline' data sets.
endline_treat_base <- endline_treat %>%
  left_join(baseline_controls, by =
              c("hhid", "group_id", "hhnomembers"))

## dropping the observations which are not in baseline survey.

only_in_baseline <- sum(is.na(endline_treat_base$base))

endline_treat_base <- endline_treat_base %>% filter(base == 1)

## writing the complete dataset
write_dta(endline_treat_base, "endline_treat_base.dta")
```

Those observations which do not match can be dropped. 360 households are not in baseline survey but are in endline survey. Hence controls are not needed for these households.

2. Analysis

a.) Hypothesis

A possible outcome of this intervention could be the increase in household formal borrowing.

H0: No increase in total formal borrowed amount.
Ha: an increase in total formal borrowed amount

Justification: Due to the expansion of local bank infrastructure in villages, availability of financial services would improve. Reliance on informal channel of borrowing would decrease. Another prior could be an increase in household income due to increased availability of financial services, that too at favourable terms.

b.) t-test of baseline household variables

Head of households' years of education or 'educyears_hoh'.

For households with their head of household having more years of education, it is likely that they will have better understanding of formal banking. If the treatment group differs from control group in this regard then the impact of financial expansion is going to be biased. For p-value = 0.9732, at 95% confidence interval, we fail to reject the null hypothesis that the difference of mean of 'education of head of household' is zero. The households are statistically similar in treatment and control groups in terms of 'education of head of household'.

```
## testing the significance of the difference

t.test(educyears_hoh ~ treated, data = endline_treat_base)
```

```
##
##  Welch Two Sample t-test
##
## data:  educyears_hoh by treated
## t = 0.033597, df = 3796.2, p-value = 0.9732
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.3020383  0.3125701
## sample estimates:
## mean in group 0 mean in group 1
##        7.438202        7.432936
```

No of household members over the age of 18 or 'hhnomembers_above18'.

If the number of members over the age of 18 differ across the treatment and control group then this could affect our outcomes. If treatment group households have less number of household members over the age of 18 compared to the control group then we can expect a low take up even when we expand the financial services. This comes from the fact that most financial services (including loans) are provided to adults alone. For p-value = 0.3962, at 95% confidence interval, we fail to reject the null hypothesis that the difference of mean of 'No of household members over the age of 18' is zero. The households are statistically similar in treatment and control groups in terms of 'No of household members over the age of 18'.

```
## testing the significance of the difference

t.test(hhnomembers_above18 ~ treated, data = endline_treat_base)
```

```
##
##  Welch Two Sample t-test
##
## data:  hhnomembers_above18 by treated
## t = 0.84846, df = 3777.4, p-value = 0.3962
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.05003742  0.12638627
## sample estimates:
## mean in group 0 mean in group 1
##        3.172820        3.134645
```

Caste variables

There is a vast literature that suggests that caste plays an important role in availability[2] and accessibility of financial services[3]. We also observe that the proportion of 'most backward caste' is lower in control group whereas the proportion of 'Backward caste' is lower in treatment group and vice versa. It is likely that this reversed proportion in control and treatment groups for these two classifications of backward communities could balance out and make control and treatment groups as a whole similar in terms of proportion of disadvantaged communities. Similarity of control and treatment group is required for the results of intervention to be valid. Dissimilarity in terms of representation of lower caste people in control and treatment group could put a question mark at the validity of our results.

If the family is of Forward Caste or 'hhcaste_fc'.

For p-value = 0.2495, at 95% confidence interval, we fail to reject the null hypothesis that the difference of mean 'whether or not the household is of forward caste' is zero. The proportion of forward caste households is statistically similar in treatment and control groups.

```
## testing the significance of the difference

t.test(hhcaste_fc ~ treated, data = endline_treat_base)
```

```
##
##  Welch Two Sample t-test
##
## data:  hhcaste_fc by treated
## t = -1.1516, df = 3650, p-value = 0.2495
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.007972310  0.002072292
## sample estimates:
## mean in group 0 mean in group 1
##     0.004817987     0.007767996
```

[2]Karthick, V., and S. Madheswaran. "Access to Formal Credit in the Indian Agriculture: Does Caste matter." Journal of Social Inclusion Studies 4.2 (2018): 169-195.

[3]Kumar, Sunil Mitra. "Does access to formal agricultural credit depend on caste?." World Development 43 (2013): 315-328.

Proportion of Backward Caste or 'hhcaste_bc'.

For p-value = 0.005375, at 95% confidence interval, we reject the null hypothesis that the difference of proportion of 'Backward caste' is zero. The households are not statistically similar in treatment and control groups in this regard.

```
## testing the significance of the difference

t.test(hhcaste_bc ~ treated, data = endline_treat_base)
```

```
##
##  Welch Two Sample t-test
##
## data:  hhcaste_bc by treated
## t = 2.7853, df = 3786.5, p-value = 0.005375
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.01307492 0.07524459
## sample estimates:
## mean in group 0 mean in group 1
##       0.4170236       0.3728638
```

Proportion of Most Backward Caste or 'hhcaste_mbc'.

For p-value = 0.002957, at 95% confidence interval, we reject the null hypothesis that the difference of proportion of 'Most backward caste' is zero. The households are not statistically similar in treatment and control groups in this regard.

```
## testing the significance of the difference

t.test(hhcaste_mbc ~ treated, data = endline_treat_base)
```

```
##
##  Welch Two Sample t-test
##
## data:  hhcaste_mbc by treated
## t = -2.9741, df = 3797, p-value = 0.002957
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.07570132 -0.01554851
## sample estimates:
## mean in group 0 mean in group 1
##       0.3158458       0.3614707
```

c.) OLS regression.

Regression of household income on the treatment dummy.
We include pair fixed effect because certain attributes of the service area pair although could remain the same over time but these attributes could differ across other pairs. Basic OLS regression model does not consider heterogeneity across pairs.

After the inclusion of pair fixed effect, 'treated' variable becomes significant. This is because all other variation in 'hhinc' has been accounted for with the inclusion of 'pair_id' variable.

Basic OLS regression

```
income_ols_lm_mod <- lm(hhinc ~ treated, data = endline_treat_base)
summary(income_ols_lm_mod)
```

```
##
## Call:
## lm(formula = hhinc ~ treated, data = endline_treat_base)
##
## Residuals:
## LABEL: household income
## VALUES:
## -10888, -7888, -4094, 895, 204096
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10094.4      409.5  24.652   <2e-16 ***
## treated        793.7      574.3   1.382    0.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17690 on 3794 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.0005032,  Adjusted R-squared:  0.0002398
## F-statistic:  1.91 on 1 and 3794 DF,  p-value: 0.167
```

FE OLS regression

```
income_fe_lm_mod <- lm(hhinc ~ treated + pair_id - 1, data = endline_treat_base)
summary(income_fe_lm_mod)
```

```
##
## Call:
## lm(formula = hhinc ~ treated + pair_id - 1, data = endline_treat_base)
##
## Residuals:
## LABEL: household income
## VALUES:
## -16989, -6664, -1857, 3817, 208585
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## treated  3945.57     525.36    7.51 7.32e-14 ***
## pair_id   237.16      11.21   21.15  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18020 on 3794 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.2329, Adjusted R-squared:  0.2325
## F-statistic:   576 on 2 and 3794 DF,  p-value: < 2.2e-16
```

clustered standard error.

```
### I could not compute 'clustered standard error' because I did not know the right command for this.
```

d.) OLS regression with 'log income' variable.

Specification 1

```
## creating  a variable  log of household income as 'log_hhinc'.
endline_treat_base <- endline_treat_base %>% mutate(log_hhinc = log(hhinc))

## We specify 'na.action' to exclude all the NAs and such values. We rerun the fixed effect model with
logincome_fe_lm_mod <- lm(log_hhinc ~ treated +
                            pair_id - 1, na.action(na.exclude(endline_treat_base)),
                          data = endline_treat_base)
summary(logincome_fe_lm_mod)
```

```
##
## Call:
## lm(formula = log_hhinc ~ treated + pair_id - 1, data = endline_treat_base,
##     subset = na.action(na.exclude(endline_treat_base)))
##
## Residuals:
## LABEL: household income
## VALUES:
## 1.91120048022294, -0.0661377589527211, -0.359855017493592, 1.7604085187065, 0.00926184816025627, -1.8
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## treated  2.10418    1.01779   2.067   0.0935 .
## pair_id  0.17074    0.01515  11.272  9.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.46 on 5 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.9787, Adjusted R-squared:  0.9702
## F-statistic: 114.8 on 2 and 5 DF,  p-value: 6.63e-05
```

Specification 2

```
## We filter the original data by selecting only those values of 'loh_hhinc' which is greater than 0.
endline_treat_base_2 <- endline_treat_base %>% filter(log_hhinc > 0)

## We rerun the original fixed effect model with log of household income and filtered data set.
logincome_fe_lm_mod_2 <- lm(log_hhinc ~ treated +
                              pair_id - 1,
```

```
                            data = endline_treat_base_2)
summary(logincome_fe_lm_mod_2)
```

```
##
## Call:
## lm(formula = log_hhinc ~ treated + pair_id - 1, data = endline_treat_base_2)
##
## Residuals:
## LABEL: household income
## VALUES:
## -8.5761, -1.0049, 1.5103, 4.1007, 10.8155
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## treated 3.292835   0.115102   28.61   <2e-16 ***
## pair_id 0.186574   0.002456   75.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.833 on 3580 degrees of freedom
## Multiple R-squared:  0.8107, Adjusted R-squared:  0.8106
## F-statistic:  7664 on 2 and 3580 DF,  p-value: < 2.2e-16
```

Interpretation and comparison.

In specification 1 the variable 'treated' comes insignificant now. Although 'treated' becomes significant in specification 2 but this could be due to the fact that the data set is now much smaller.

e.) OLS regression with Household controls.

Variables such as 'readwrite_hoh', 'educyears_hoh', 'higheduc_hoh', and 'noclasspassed_hoh' are highly correlated among each other. Only 'educyears_hoh' is taken into regression for only it could act as a good proxy for the rest. Other variables included are 'gender_hoh', because it is likely that female headed household could be poorer[4], 'hhnomembers_above18', because more adult household members could translate into more working hands and hence more income, 'hhcaste_bc' and 'hhcaste_sc_st', because there are several disadvantages associated with lower caste. Additionally, we observe that variables 'educyears_hoh', 'hhnomembers_above18', and 'hhcaste_sc_st' are significant.
'treated' remains insignificant with the inclusion of controls.

```
income_fe_lm_mod_hhc <- lm(hhinc ~ treated + gender_hoh + educyears_hoh +
                            hhnomembers_above18 + hhcaste_bc + hhcaste_sc_st +
                            pair_id - 1, data = endline_treat_base)
summary(income_fe_lm_mod_hhc)
```

```
##
## Call:
## lm(formula = hhinc ~ treated + gender_hoh + educyears_hoh + hhnomembers_above18 +
##     hhcaste_bc + hhcaste_sc_st + pair_id - 1, data = endline_treat_base)
##
```

[4]Rajaram, Ramaprasad. "Female-headed households and poverty: evidence from the National Family Health Survey." University of Georgia, USA (2009): 132-137

```
## Residuals:
## LABEL: household income
## VALUES:
## -22981, -7128, -3238, 1185, 204688
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## treated                737.22     536.99   1.373  0.16987
## gender_hoh             256.81     654.27   0.393  0.69470
## educyears_hoh          431.89      58.19   7.423 1.41e-13 ***
## hhnomembers_above18   1906.85     176.30  10.816  < 2e-16 ***
## hhcaste_bc              12.26     623.04   0.020  0.98430
## hhcaste_sc_st        -2242.77     690.97  -3.246  0.00118 **
## pair_id                 46.32      15.73   2.945  0.00325 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17260 on 3788 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.2972, Adjusted R-squared:  0.2959
## F-statistic: 228.8 on 7 and 3788 DF,  p-value: < 2.2e-16
```

```
## Using stargazer package to export regression table.

stargazer(income_fe_lm_mod_hhc, type = "text",
          title = "impact of treatment on household income, an OLS regression: pair fixed effect with h
          out = "income_fe_lm_mod_hhc.txt")
```

```
##
## impact of treatment on household income, an OLS regression: pair fixed effect with household level c
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                hhinc
## ------------------------------------------------
## treated                        737.223
##                               (536.994)
##
## gender_hoh                     256.810
##                               (654.267)
##
## educyears_hoh                431.895***
##                               (58.185)
##
## hhnomembers_above18         1,906.854***
##                               (176.297)
##
## hhcaste_bc                     12.258
##                               (623.044)
##
## hhcaste_sc_st               -2,242.772***
##                               (690.965)
##
## pair_id                       46.319***
```

```
##                               (15.730)
##
## ------------------------------------------------
## Observations                    3,795
## R2                              0.297
## Adjusted R2                     0.296
## Residual Std. Error   17,258.510 (df = 3788)
## F Statistic          228.833*** (df = 7; 3788)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

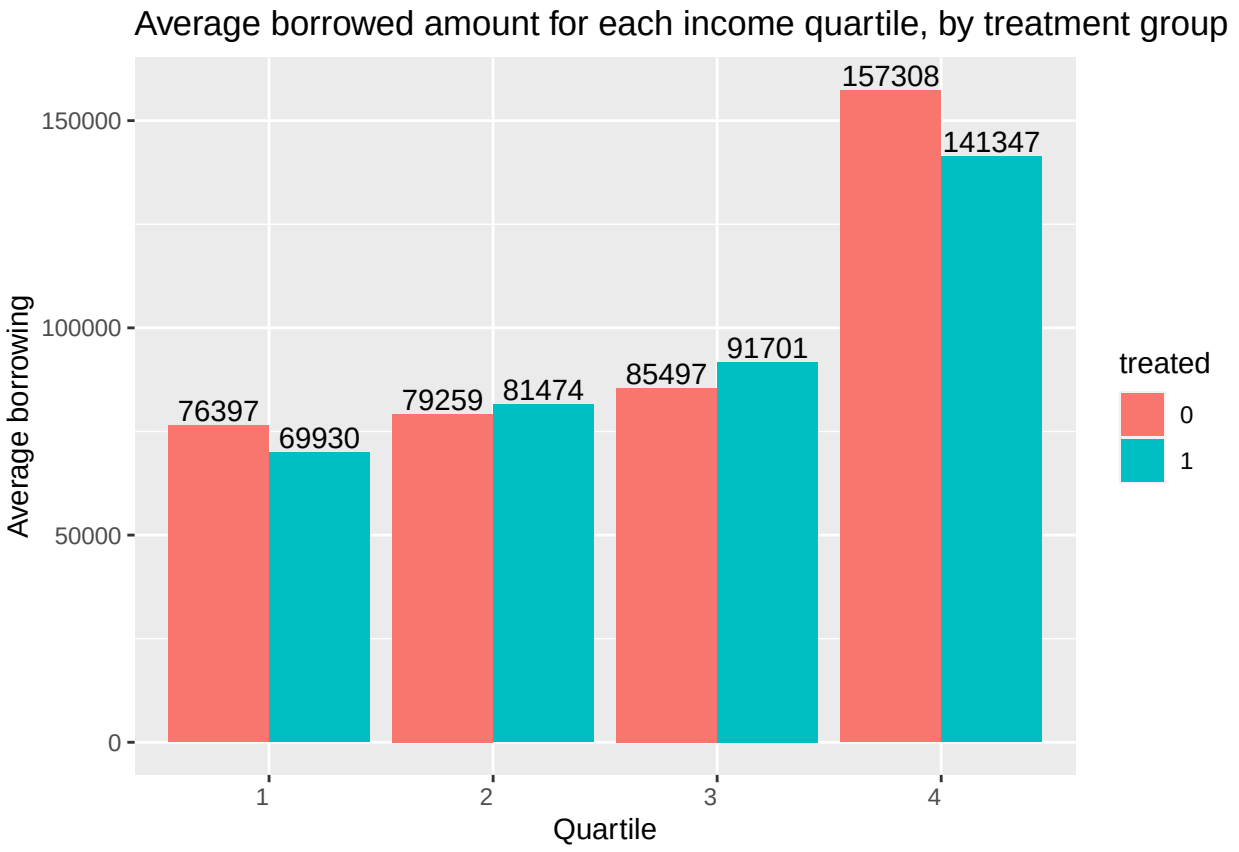
f.) Bar chart: borrowed amount for each income quartile, by treatment group.

```
## Calculate the 1st quartile, median, and the 3rd quartile.


endline_treat_base$quartile <- ntile(endline_treat_base$hhinc, 4)

endline_treat_base$quartile <- as.factor(
endline_treat_base$quartile)

borrowing_data <- endline_treat_base %>%
  select(quartile, tot_borrw, treated) %>%
  filter(!is.na(tot_borrw)) %>%
  filter(!is.na(quartile)) %>%
  group_by(quartile, treated) %>%
  summarise(avg_borr = mean(tot_borrw,))
```

```
## `summarise()` has grouped output by 'quartile'. You can override using the `.groups` argument.
```

```
borrowing_data$avg_borr <- round(borrowing_data$avg_borr)

borrowing_data$treated <- as.factor(borrowing_data$treated)
```

```
avg_borr_plot <- ggplot(borrowing_data,
       aes(x = quartile, y = avg_borr,
           fill = treated)) +
  geom_bar(stat = "identity", position=position_dodge())+
  ggtitle(label = "Average borrowed amount for each income quartile, by treatment group") +
  labs(x = "Quartile", y = "Average borrowing") +
theme(axis.text.x = element_text(angle = 0, hjust = 1)) +
geom_text(aes(label = avg_borr), vjust = -0.2, size = 4,
position = position_dodge(0.9))
avg_borr_plot
```




```
ggsave("avg_borr_plot.png", width = 9, height = 5, units = "in", dpi = 300)
```

Experience with R.

I became familiar with R during my masters. I have used R to prepare assignments and term papers. Thereafter I have used R in my current work as an RA.