Homework 2: Evaluating Randomized Experiments

Neeraj Sharma

05/02/2020

Question 1: Estimate the treatment effect and the associated standard error on the raw data given to you, assuming no issues with the data.

By definition, the Average Treatment Effect is defined to be $\frac{1}{N}\sum_i y_1(i) - y_0(i)$ where $y_1(i)$ and $y_0(i)$ are the values of the outcome variable (in this case hours exercised) in each treatment scenario. Unfortunately, it is often impractical to utilize this straightforward approach. This formula presumes that one can quantify the outcome of a given individual when treatment is both given and withheld. However, because each individual can only be slotted into one category, this approach cannot be perfectly achieved.

Thus, a random experiment with a control and treatment group can be conducted to smooth out differences among populations in order to isolate the treatment effect specifically. With large enough sample groups, the difference between the mean of the outcome of the treatment group and mean of the outcome of the control group yields the Average Treatment Effect.

Here are the relevant summary statistics of the data set without any modification.

Table 1: Summary of Hours Exercised, No Data Cleaning

Treatment	Count	Mean	St Dev	St Err
0	500	20.952	84.02863	3.757875
1	500	27.546	104.36699	4.667434

The mean hours exercised for individuals in the treatment group is 27.546 and the mean hours exercised for individuals in the non-treatment group is 20.952. Assuming randomness was properly implemented in this study, the difference between these two numbers will be the Average Treatment Effect of the treatment based on the analysis I provide above. Thus, the Average Treatment Effect is **6.594** with a standard error of **0.9096**.

Question 2: It's always a good idea to check a data set for errors. Clean this data set as you think appropriate. As an answer to this question, note all the kinds of changes you made to the data, a few words explaining your reason for the change, and which observations you changed (noting the observation number included as a variable in the data set for identification purposes). If it is totally obvious which observations you changed and there are a large number in the category (e.g. if you decided to drop all White study participants), you can just note what you did for that change ("I dropped all white participants") and explain why.

I noticed several types of data errors and compiled a representative sample of troubled observations here.

Table 2: Representative sample of unclean observations

subject_id	hours	treatment	community_center	female	age	bmi	education	race_ethnicity
1	2	0	Woodlawn	0	44	28.6634560	high school	BLACK
9	14	0	WOODLAWN	0	19	29.3965780	high school	black
16	180	0	hyde park	1	32	26.9350070	higher degree	BLACK

subject_id	hours	treatment	community_center	female	age	bmi	education	race_ethnicity
26	4	0	hyde park	0	-99	33.4841000	higher degree	black
31	13	1	Woodlawn	female	22	22.7878760	higher degree	BLACK
52	1	0	hyde park	male	51	35.4018250	higher degree	white
65	11	1	Woodlawn	0	20	31.7927250	high school	NA
100	7	0	Hyd Park	1	39	34.9566570	higher degree	white
103	8	1	Woodlawn	1	37	0.2565794	high school	BLACK

Given these errors, I perform the following modifying operations to clean the data set. See appendices for the specific code I use to accomplish these modifications.

Variable	Description of Modification
subject_id	No change
hours	Hours over 60 are minutes; reformatted to be hours.
treatment	No change.
community_center	All variant spellings recoded to "Woodlawn" and "Hyde Park."
female	0/1/male/female recoded uniformly as 0/1.
age	-99 means missing age data; recoded to be NA.
bmi	BMIs of less than 1 have undetermined error; recoded to be NA.*
education	No change.
race_ethnicity	Race/Ethnicity "BLACK" capitalization recoded to "Black."
changed?	`changed?` notes and counts changes that occurred in cleaning an observation.

*The source of error on BMI is very difficult to determine. There are multiple hypotheses I have for the questionable data. It could be an error in data entry where the decimal was misplaced in which case multiplying by 100 would resolve the issue. It could be that the data was entered in a different unit than normal (pounds vs kilograms or meters vs feet) and the application of the BMI formula resulted in very low numbers. It is impossible to certify which error occurred, if any, so the safest option is to ignore these 26 rows. This is a small enough number that it will not significantly harm my interpretation.

Question 3: With your cleaned data set, re-estimate the treatment effect and estimated standard error, assuming the randomization worked fine.

Table 4: Summary of Hours Exercised, Data Cleaned

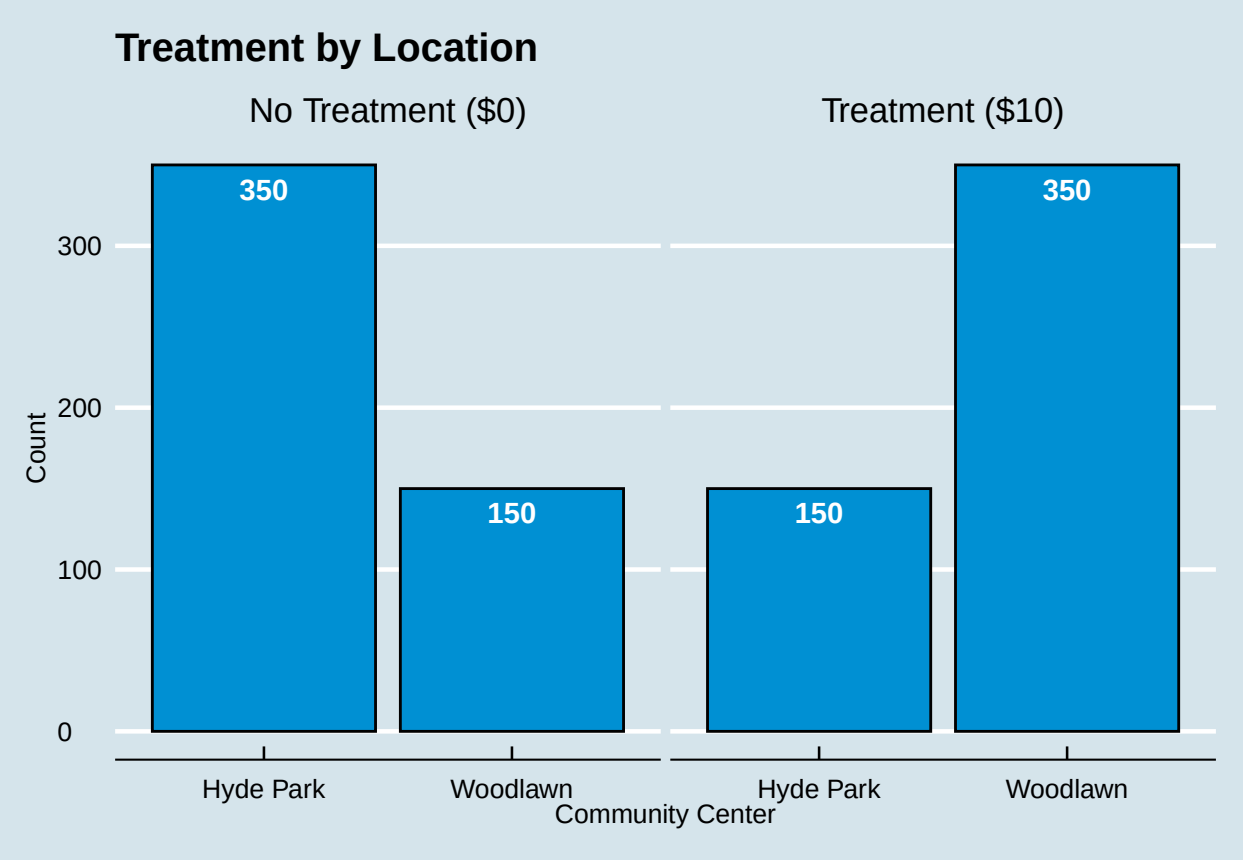
Treatment	Count	Mean	St Dev	St Err
0	500	5.612	3.387093	0.1514754
1	500	7.486	3.669432	0.1641020

The mean number of hours exercised in both the treatment and control groups has fallen dramatically upon cleaning the data. Numerous entries were coded in minutes instead of hours and those observations were dragging the values up significantly. Those errors have since been corrected. Thus, the Average Treatment Effect upon cleaning the data yet assuming proper randomization is is **1.874** with a standard error of **0.0126**.

Question 4: Evaluate whether the randomization appears legitimate. If not, what is your evidence? (Hint: something went wrong.)

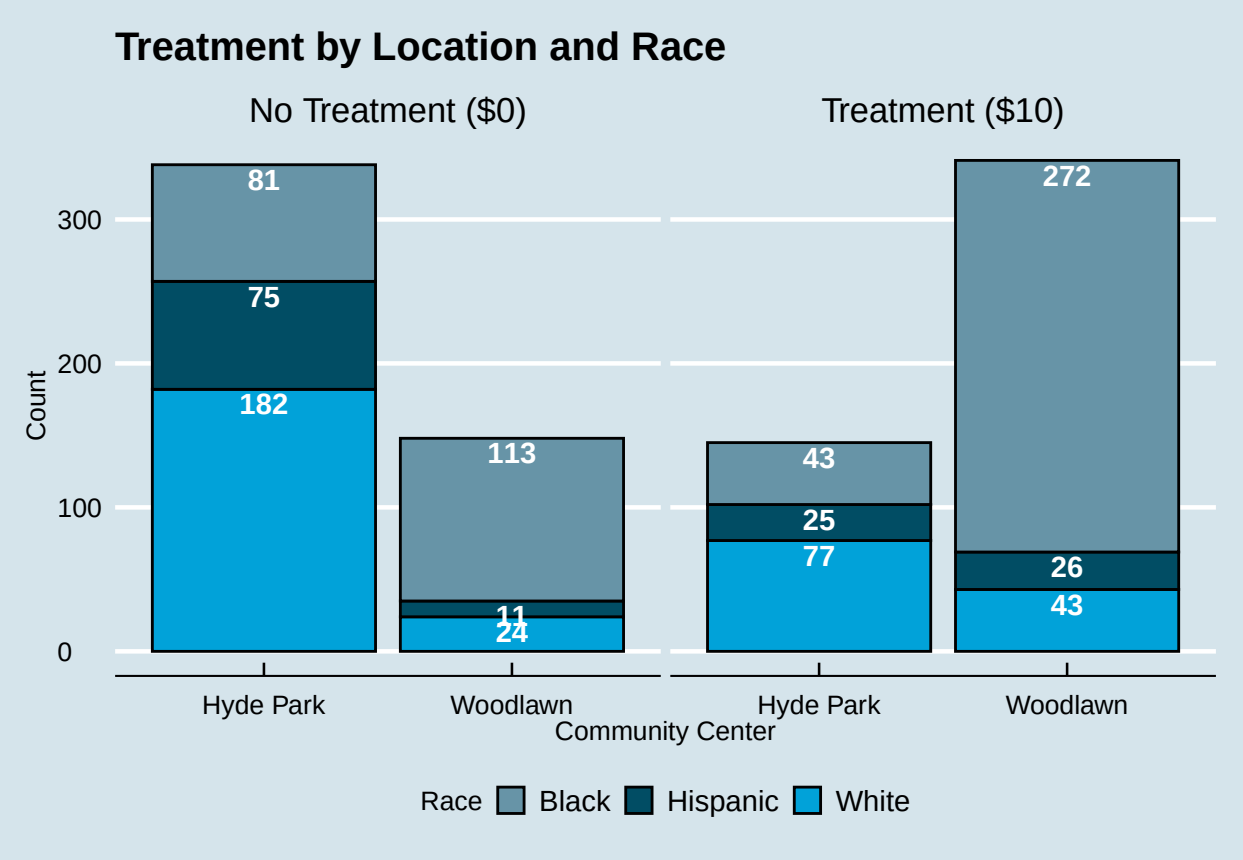
An important piece of insight I gained from Eric's office hours were that given a sufficiently large sample size, perfect random assignment means by definition that one could randomly subdivide the data into equally

sized groups of people and the distribution of the covariates should be the same. Thus, if I were to create histograms of some of the specified covariates in this data set like BMI, age, race, or community, I should observe similar distributions and summary statistics to justify the representatives and randomness of the allocation of the sample groups. Eric noted that on this problem set specifically is straightforward enough with regards to the identification of systematic differences that imply a non-random allocation that full t-test would not be necessary.

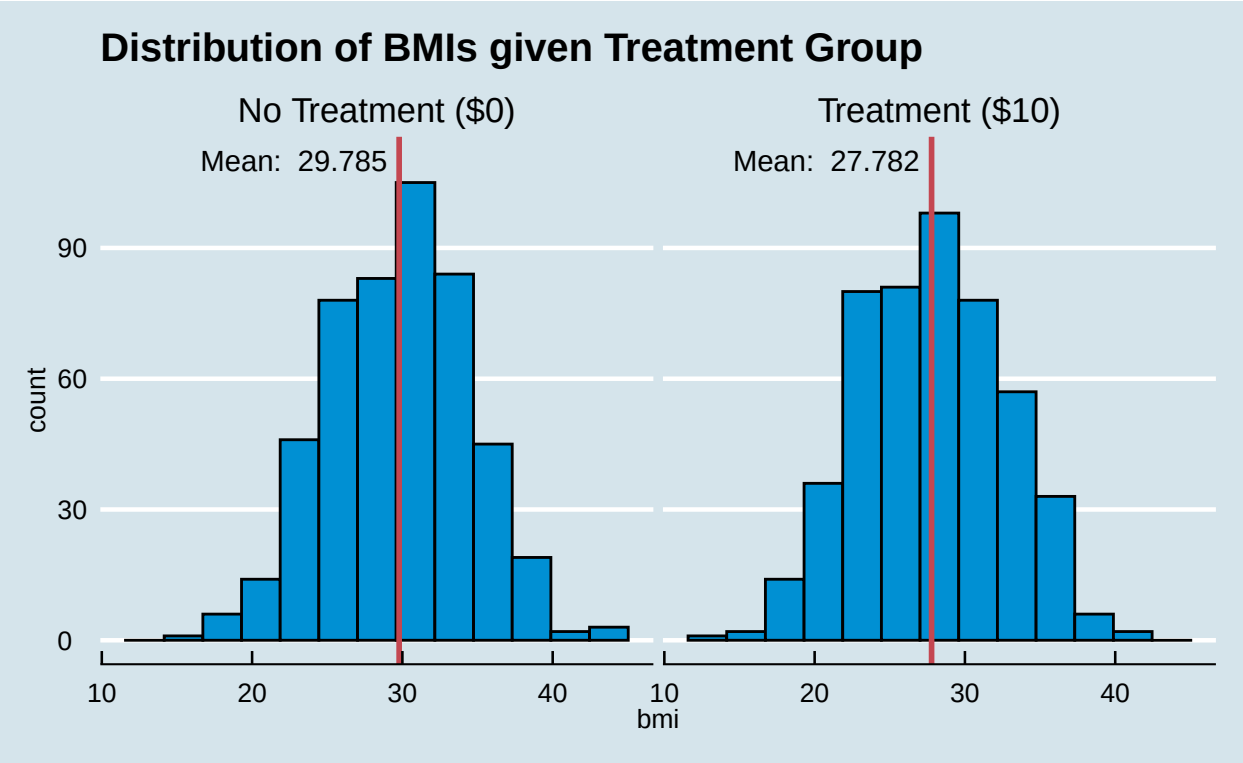


This plot clearly shows a bias in the group allocations based on survey site. Were this data to be randomly created, we would expect to see an even distribution of individuals in the treatment and non-treatment groups across Woodlawn and Hyde Park. Instead, we find a disproportionate amount of the control group was assigned at Hyde Park and conversely, a disproportionate number of people in the treatment group were assigned in Woodlawn. A random experiment would not exhibit this property.

Further analysis on other covariates supports the conclusion that race is not controlled between treatment groups.



The third and final variable I perform significant analysis on to understand the pattern of randomness is BMI.

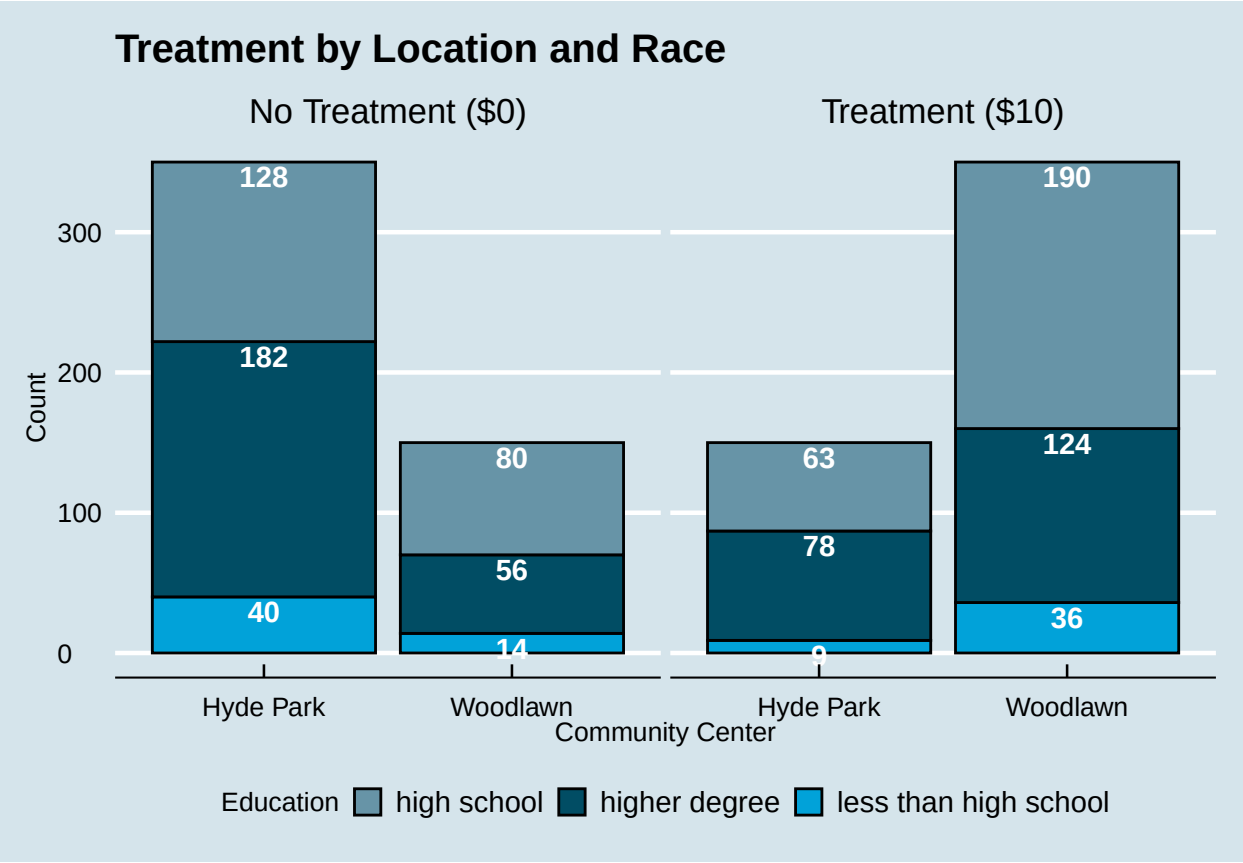


There is an apparent difference between the mean BMI in the treatment and non-treatment group, but it is unclear if it is significant enough to warrant inclusion as a source of failed randomization. A simple T-Test can shed more light.

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	alternative
2.003217	29.78517	27.78196	6.527694	1e-10	971.5728	1.400993	2.60544	two.sided

The T-Test reveals that there is a statistically significant difference between the means of the two sample groups given the p-value of 1e-10.

Plots of other covariates imply that those variables were properly randomized. For instance, the plot for education visually provides evidence for an even distribution of education levels throughout the community centers and test groups.



For brevity, I only provide summary statistics to articulate this point further:

Table 6: Summary of random covariates

treatment	Mean Female	Mean Age	less than high school	high school	higher degree
No Treatment ($0)	0.636	32.86141	54	208	238
Treatment ($10)	0.582	31.68421	45	253	202

Question 5: Offer your best hypothesis/hypotheses as to what went wrong with the randomization? What evidence do you have to support your hypothesis(es)? For each of these hypotheses, describe your best strategy for estimating a plausible treatment effect, in spite of the bad randomization. (But don't actually estimate that treatment effect.)

My hypothesis is that Justin was trigger happy with his treatment assignments because he was stationed at the Woodlawn community center which is over represented in the treatment group. Question 4 provides a hint indicating that something "went wrong" in the data collection process. This implies that there was a material mistake in the collection separate from random variance in the population that might skew covariates in one direction or the other. The number of observations taken from each sample site is something that is in Eric and Justin's control, so I believe that this error is the thing that "went wrong" and is the source of the error in randomization.

The evidence I have collected to support this hypothesis is the chart I produce above as well as the following t-test which rejects the two-tailed null hypothesis that the means are equal.

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	alternative
-0.394269	0.2983539	0.692623	-13.37631	0	971.9819	-0.4521113	-0.3364267	two.sided

Furthermore, I hypothesize that the differences in race and BMI are also side effects of this incorrect sampling. As far as race goes, Hyde Park has the University of Chicago which dramatically alters the racial mixture of the south-side neighborhood. Most University affiliates are white, and most live in Hyde Park, so it makes sense that the Hyde Park neighborhood will have a significantly higher number of white individuals sampled compared to Woodlawn.

Secondly, I have absolutely zero evidence to back this up, but my gut is that Hyde Park has a lower average BMI than Woodlawn due to population differences in terms of primarely income. I have not explored this avenue however, so I would like to place a significant caveat on this point. No matter what, both race and BMI are attributes that are intrinsic to the neighborhoods but are also important to control for.

My strategy to control for the community center, race, and BMI data is to simply do a multiple linear regression across those, plus the independent treatment variable. I have coded dummy variables for race and community center as those are the primary categoricals I will be analyzing.

Question 6: Given your answer to question five come up with your best estimate of the true treatment effect in the experiment as well as its standard error.

Degrees of Freedom	Residual Sum of Squareds	delta DF	Sum of Squares	F Statistic	Pr (> F)
972	12167.934	NA	NA	NA	NA
971	10989.362	1	1178.5725	122.44650	0
969	9818.652	2	1170.7098	60.81480	0
968	9317.198	1	501.4543	52.09804	0

Comparing the models confirms that a model with treatment, community center, race, and bmi is meaningfully more accurate than any other given the randomness and treatment type, I proceed with that as my model.

Table 9: Summary of Hours Exercised, Pure Treatment Effect

Treatment	Count	Mean	St Dev	St Err
0	486	5.606996	1.753335	0.0795329

Treatment	Count	Mean	St Dev	St Err
1	488	7.502049	1.670961	0.0756408

After all this, the average treatment effect calculated after adjusting for errors in randomness and corresponding covariates is **1.8951** and the standard error on that treatment effect is **0.0039**.

Appendices

Code to clean data in question 2

```
# Clean the data to impose uniformity upon the variable encoding.
clean_data <- raw %>%
  mutate(`changed?` = seq(1, 1000) * 0) %>%
  # Fixing messed up hours readings. They start at 60 and upwards and that's 1 hr so I
  # fix based on that.
  mutate(`changed?` = if_else(hours >= 60, `changed?` + 1, `changed?`),
         hours = if_else(hours >= 60, hours / 60, hours)) %>%
  mutate(`changed?` = if_else(community_center %in% c("WOODLAWN",
                                                      "hyde park",
                                                      "woodlawn",
                                                      "Hyd Park",
                                                      "HYDE PARK",
                                                      "Hyde_Park",
                                                      "HYDE_PARK",
                                                      "hyde_park"),
                              `changed?` + 1, `changed?`),
         community_center = if_else(community_center %in% c("WOODLAWN", "woodlawn"),
                                    "Woodlawn", community_center),
         community_center = if_else(community_center %in% c("hyde park",
                                                            "Hyd Park",
                                                            "HYDE PARK",
                                                            "HYDE_PARK",
                                                            "hyde_park",
                                                            "Hyde_Park"),
                                    "Hyde Park", community_center)) %>%
  # Recoding female variable to be factor categorical variable from 0/1/male/female.
  mutate(`changed?` = if_else(female %in% c("female", "male"),
                              `changed?` + 1, `changed?`),
         female = as.double(if_else(female == "female",
                                    "1", if_else(female == "male", "0", female)))) %>%
  # -99 is missing age data so I reincode it at missing age data.
  # https://cran.r-project.org/web/packages/naniar/vignettes/replace-with-na.html
  mutate(`changed?` = if_else(age == -99, `changed?` + 1, `changed?`),
         age = na_if(age, -99)) %>%
  # Currently I have removed improper values, but I could also justify multiplying by 10.
  mutate(`changed?` = if_else(bmi < 1, `changed?` + 1, `changed?`),
         bmi = ifelse(bmi < 1, NA, bmi)) %>%
  # Fix all BLACK observations to normal capitalization structure.
  mutate(`changed?` = if_else(is.na(race_ethnicity),
                              `changed?`, if_else(race_ethnicity == "BLACK",
                                                  `changed?` + 1, `changed?`)),
    race_ethnicity = if_else(is.na(race_ethnicity),
                             race_ethnicity, str_to_sentence(race_ethnicity))) # %>%
```

```
# mutate(race_ethnicity = factor(race_ethnicity, c("Black", "Hispanic", "White"))) %>%
# mutate(treatment = factor(treatment, labels = c("No Treatment ($0)", "Treatment ($10)"))) %>%
# mutate(community_center = factor(community_center, labels = c("Woodlawn", "Hyde Park"))) %>%
# mutate(education = factor(education, levels = c("less than high school",
#                                                 "high school",
#                                                 "higher degree")))
```