Data Exploration: Symbolic Politics

Yao Yu

October 21, 2021

In this Data Exploration assignment we will explore Reny and Newman's (2021) finding that opinions towards the police and about the level of discrimination faced by Black Americans were impacted by the spread of protests in the wake of the killing of George Floyd. You will recreate, present, and assess those claims as well as creating your own regression models to test which attitudes change and when.

If you have a question about any part of this assignment, please ask! Note that the actionable part of each question is **bolded**.

Opinion Mobilization: The George Floyd Protests

Data Details:

- File Name: `RN_2001_data.RData`
- Source: These data are from Reny and Newman (2021).

Variable Name	Variable Description
`race_ethnicity`	Race or ethnicity. Levels labelled in data: 1-White, 2-Black or AfAm, 3-American Indian or Alaskan Native, 4 through 14- Asian or Pacific Islander (details in labels), and 15-Some other race
`hispanic`	Of Hispanic, Latino, or Spanish origin. Levels labelled in data: 1-Not Hispanic, 2-15 Hispanic of various origins
`day_running`	Day relative to onset of George Floyd protests (day 0)
`age`	Respondent's age
`female`	Binary indicator variable: 1 if respondent female, 0 otherwise
`college`	Binary indicator variable: 1 if respondent attended college, 0 otherwise
`household_income`	Household pre-tax income ranging from 1 (less than $15,000) to 24 (more than $250,000). Details for other levels in labels.
`pid7`	Party identification on a seven point scale with strong, weak, lean: 1-Strong Democrat to 7-Strong Republican with 4-Independent.
`ideo5`	Ideological self placement: 1-Very liberal, 2-Liberal, 3-Moderate, 4-Conservative, 5-Very Conservative
`vote_clinton`	Indicator variable for whether the respondent said they voted for Clinton in the 2016 presidential election
`group_favorability_the_police`	Favorability towards the police: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable
`discrimination_blacks`	Perceptions of the level of discrimination in US faced by Blacks: 1-None at all, 2-A little, 3-A moderate amount, 4-A lot, 5-A great deal
`date`	The date the respondent took the survey

Variable Name	Variable Description
group_fav_white_black	The difference in respondents favorability towards Blacks subtracted from their favorability towards whites (each on four point scale). Ranges from -3 to 3.
racial_attitudes_generations	Agreement with the statement that generations of slavery and discrimination have made it difficult for Blacks to work their way out of the lower class: 1-Strongly Agree to 5-Strongly Disagree
interest	Degree to which respondent claims to follow politics: 1-Most of the time, 2-Some of the time, 3-Only now and then, 4-Hardly at all
group_favorability_jews	Favorability towards Jews: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable
group_favorability_whites	Favorability towards whites: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable
group_favorability_evangelicals	Favorability towards evangelicals: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable
group_favorability_socialists	Favorability towards socialists: 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, 4-Very unfavorable
protest	Indicator variable if survey respondent lived in area that would at any point have a BLM protest in the wake of the killing of George Floyd
n_protests	Number of eventual BLM protests in area where resident lived

```
# load the data containing the tibble protest_df
load('RN_2001_data.RData')

#Note that the data is saved in the form of a tibble, a special table using the dplyr package that has

head(protest_df$race_ethnicity)
```

```
## <labelled<double>[6]>: What is your race? Provided by LUCID.
## [1] 6 1 1 2 1 1
##
## Labels:
##  value                              label
##      1                              White
##      2         Black, or African American
##      3   American Indian or Alaska Native
##      4               Asian (Asian Indian)
##      5                    Asian (Chinese)
##      6                   Asian (Filipino)
##      7                   Asian (Japanese)
##      8                     Asian (Korean)
##      9                  Asian (Vietnamese)
##     10                       Asian (Other)
##     11 Pacific Islander (Native Hawaiian)
##     12       Pacific Islander (Guamanian)
##     13          Pacific Islander (Samoan)
##     14           Pacific Islander (Other)
##     15                     Some other race
##    777             Not asked in this wave
```

```
head(protest_df$household_income)
```

```
## <labelled<double>[6]>: What is your current annual household income before taxes? Provided by L...
```

```
## [1] 21  8  7  1 NA  1
##
## Labels:
##  value                  label
##      1      Less than $14,999
##      2     $15,000 to $19,999
##      3     $20,000 to $24,999
##      4     $25,000 to $29,999
##      5     $30,000 to $34,999
##      6     $35,000 to $39,999
##      7     $40,000 to $44,999
##      8     $45,000 to $49,999
##      9     $50,000 to $54,999
##     10     $55,000 to $59,999
##     11     $60,000 to $64,999
##     12     $65,000 to $69,999
##     13     $70,000 to $74,999
##     14     $75,000 to $79,999
##     15     $80,000 to $84,999
##     16     $85,000 to $89,999
##     17     $90,000 to $94,999
##     18     $95,000 to $99,999
##     19   $100,000 to $124,999
##     20   $125,000 to $149,999
##     21   $150,000 to $174,999
##     22   $175,000 to $199,999
##     23   $200,000 to $249,999
##     24     $250,000 and above
##    777 Not asked in this wave
```

Question 1

As usual it is important to first examine the structure of the data. What are the two main outcome variables of interest to Reny and Newman? How were they measured and how are they coded in the data? What was the treatment? **Take a look at the data and determine which are the two outcome variables of interest. Observe the scale of each.**

The two main outcome variables of interest to Reny and Newman are `group_favorability_the_police`, the favorability towards the police, and `discrimination_blacks`, perceptions of the level of discrimination in US faced by Blacks. The measurement for `group_favorability_the_police` is 1-Very favorable, 2-Somewhat favorable, 3-Somewhat unfavorable, and 4-Very unfavorable. The measurement of `discrimination_blacks` is 1-None at all, 2-A little, 3-A moderate amount, 4-A lot, and 5-A great deal. The treatment variable here would be `protest`, an indicator variable if survey respondent lived in area that would at any point have a BLM protest in the wake of the killing of George Floyd.

Question 2

Part a

R has a special 'date' class for storing and manipulating dates as seen below. Date variables can conveniently be logically compared and arithmetically manipulated. Using the day variable find out how many days the dataset spans. **First check using the code below that the day variable is of the class 'date'. Next subtract the latest day in the sample from the first day to calculate the timespan covered by the dataset. Hint: functions like max() and min() work for date variables too!**

```
class(protest_df$day)
```

```
## [1] "Date"
```

```
difftime(max(protest_df$day), min(protest_df$day))
```

```
## Time difference of 419 days
```

Part b

On what date is the treatment said to have occurred? **Find the date for which the day_running variable is 0. Then modify the code below to add a variable to each row for whether or not the observation was before or after treatment.**

```
# Pull the date when day_running is 0
day_of_protest <- protest_df %>%
  filter(day_running == 0) %>%
  pull(day) %>%
  .[1]

print(day_of_protest)
```

```
## [1] "2020-05-28"
```

```
#Change the object to be the date of the protest spread, remember to put it in quotes if you copy/paste

protest_df_bydate <- protest_df %>%
  mutate(before = ifelse(day<as.Date(day_of_protest), 1,0))
```

Question 3

Part a

Compare the average for each outcome variable before and after the onset of the protests. Are the differences statistically significant? **Calculate the outcome variable means for before and after treatment. Conduct a test as to whether the differences in means are statistically significant. Hint: you can use either the t.test() function or difference_in_means() from the estimatr package**

```
difference_in_means(group_favorability_the_police ~ before, data = protest_df_bydate)
```

```
## Design:  Standard
##          Estimate  Std. Error   t value      Pr(>|t|)   CI Lower   CI Upper
## before -0.1547348 0.004137599 -37.39725 1.382628e-304 -0.1628445 -0.1466252
##              DF
## before 140455.2
```

```
difference_in_means(discrimination_blacks ~ before, data = protest_df_bydate)
```

```
## Design:  Standard
##          Estimate  Std. Error   t value      Pr(>|t|)   CI Lower   CI Upper
## before -0.1369509 0.004587892 -29.85051 3.030341e-195 -0.1459431 -0.1279587
##              DF
## before 158036.6
```

Comparing the average for each outcome variable before and after the onset of the protests, we can see that the differences are both statistically significant. For `group_favorability_the_police`, we see an average change in attitudes of 0.155 with a 95% confidence interval of 0.147 and 0.163. For `discrimination_blacks`, we see an average change in attitudes of 0.137 with a 95% confidence interval of 0.128 and 0.146.

Part b

It might be that the period before and after the treatment was different in ways in addition to the onset of the protests. Use the same procedure as above to check for differences between two means of a survey response measuring favorability towards a group besides the police. **Calculate the means from before and after the treatment and conduct a test of statistical significance of the difference for another measure of group favorability that was recorded in the survey (e.g. evangelicals, Jews, socialists, or whites). Is there also a substantive or statistically significant difference on that variable? Should that change our confidence in attributing the opinion changes found in part a to the George Floyd protests?**

```
difference_in_means(group_favorability_jews ~ before, data = protest_df_bydate)
```

```
## Design:  Standard
##          Estimate  Std. Error  t value     Pr(>|t|)   CI Lower   CI Upper
## before 0.02258569 0.003858487 5.853508 4.822584e-09 0.01502313 0.03014824
##              DF
## before 157960.1
```

```
difference_in_means(group_favorability_whites ~ before, data = protest_df_bydate)
```

```
## Design:  Standard
##          Estimate  Std. Error  t value     Pr(>|t|)   CI Lower   CI Upper
## before 0.03943091 0.003516989 11.21155 3.683452e-29 0.03253768 0.04632414
##              DF
## before 138860.7
```

```
difference_in_means(group_favorability_evangelicals ~ before, data = protest_df_bydate)
```

```
## Design:  Standard
##          Estimate  Std. Error  t value     Pr(>|t|)   CI Lower   CI Upper
## before 0.04435605 0.004701419 9.434608 3.982521e-21 0.03514136 0.05357074
##              DF
## before 138147.3
```

```
difference_in_means(group_favorability_socialists ~ before, data = protest_df_bydate)
```

```
## Design:  Standard
##          Estimate  Std. Error  t value     Pr(>|t|)   CI Lower   CI Upper
## before 0.04452018 0.004619561 9.637318 5.663127e-22 0.03546592 0.05357444
##              DF
## before 122462.2
```

Looking at the group favorability of other groups in the survey, we continue to find changes that are statistically significant. However, the average changes in attitudes are instead slightly more positive instead of negative. This continues to support our confidence in attributing the opinion changes found in part a to the George Floyd protests as they show that other attitudes did not have changes with large magnitudes before and after the onset of protests.

Question 4

Part a

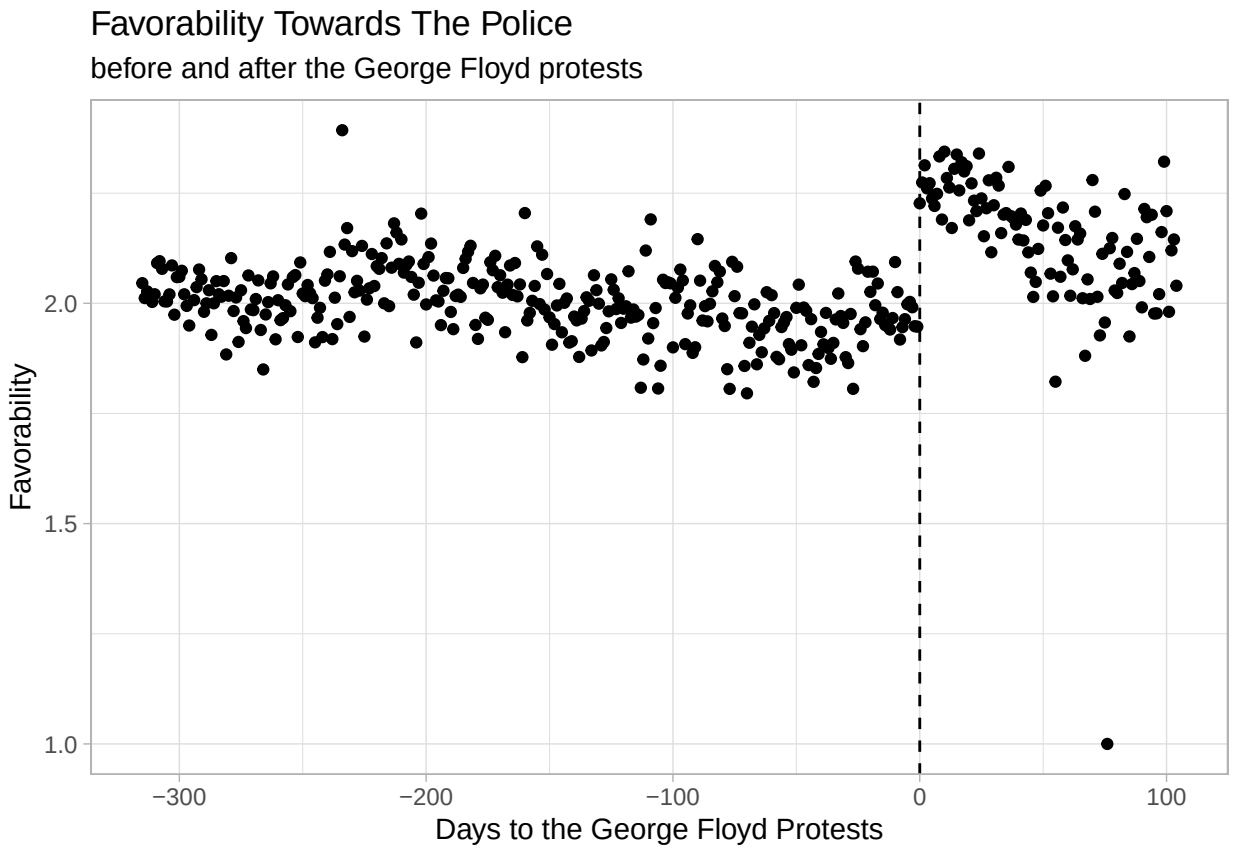
In order to create figures similar to the panels in Figure 2 in Reny and Newman (2021) we must first manipulate the data to be more usable. If we intend to graph the average of each outcome variable for each day, on what variable should we group the data using group_by? **Create a new object that is the data split out by the appropriate group and producing the average for each of the two outcome variables for each day. Also be sure to preserve an indicator for whether the observations are from before or after the spread of the protests.**

```
protest_df_byrace_ethnicity <- protest_df_bydate %>%
  mutate(race_ethnicity_grouped = case_when(
    race_ethnicity == 1 & hispanic == 1 ~ "Non-Hispanic White",
    race_ethnicity == 2 ~ "Black/African-American",
    race_ethnicity >= 4 & race_ethnicity <= 14 ~ "Asian American",
    # Unsure if this is best way of encoding Hispanics bc of overlap with other races
    hispanic != 1 ~ "Latinos",
    TRUE ~ "Other"
  ))
```

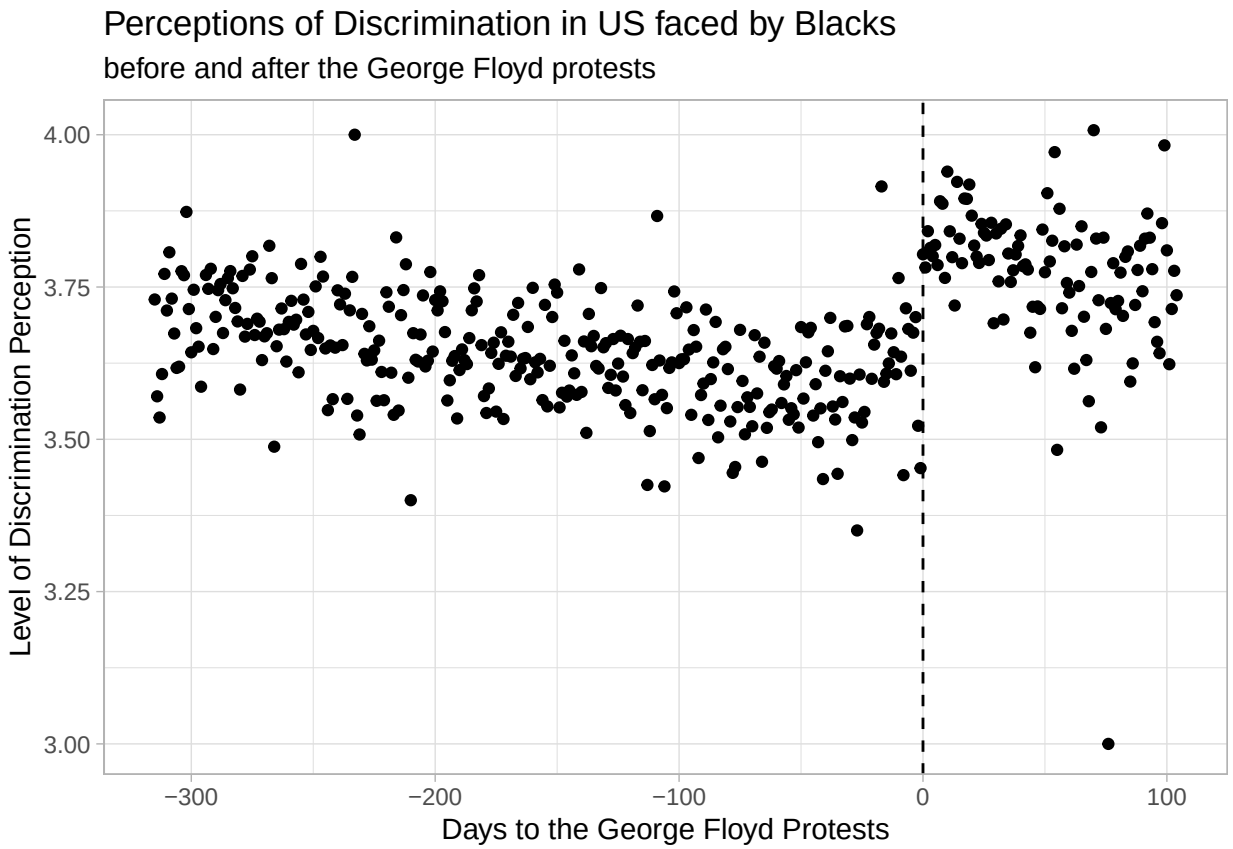
Part b

Graph the results for the entire sample. **Graph the results for the entire sample for both outcome variables by day. Include a vertical line demarcating when the protests started to spread. Does there appear to be a shift in the outcome variables from before to after the protests began to spread?**

```
# group_favorability_the_police
protest_df %>%
  drop_na(group_favorability_the_police) %>%
  group_by(day_running) %>%
  summarize(avg_group_favorability_the_police = mean(group_favorability_the_police),
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = avg_group_favorability_the_police)) +
  geom_point() +
  geom_vline(xintercept = 0, linetype = "dashed") +
  theme_light() +
  labs(
    title = "Favorability Towards The Police",
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Favorability"
  )
```



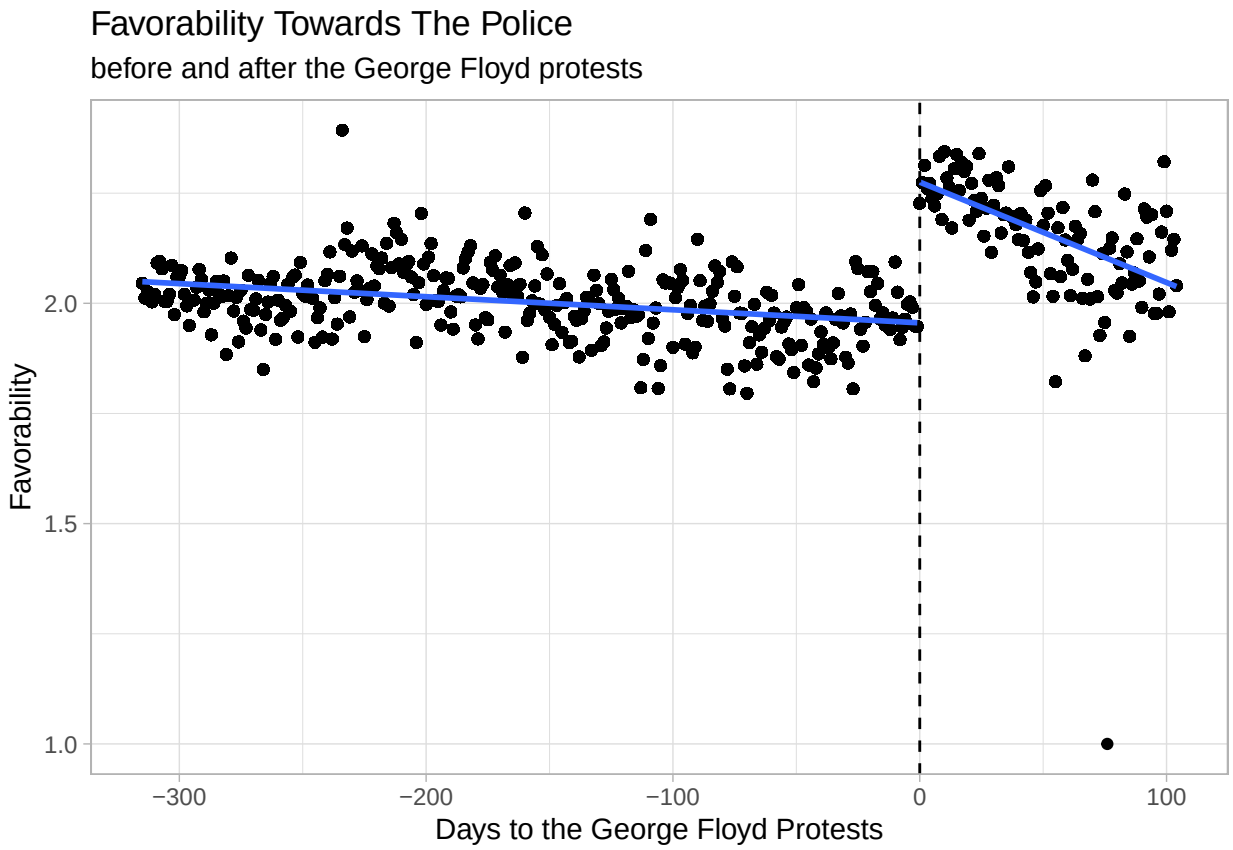

```
# discrimination_blacks
protest_df %>%
  drop_na(discrimination_blacks) %>%
  group_by(day_running) %>%
  summarize(avg_discrimination_blacks = mean(discrimination_blacks),
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = avg_discrimination_blacks)) +
  geom_point() +
  geom_vline(xintercept = 0, linetype = "dashed") +
  theme_light() +
  labs(
    title = "Perceptions of Discrimination in US faced by Blacks",
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Level of Discrimination Perception"
  )
```



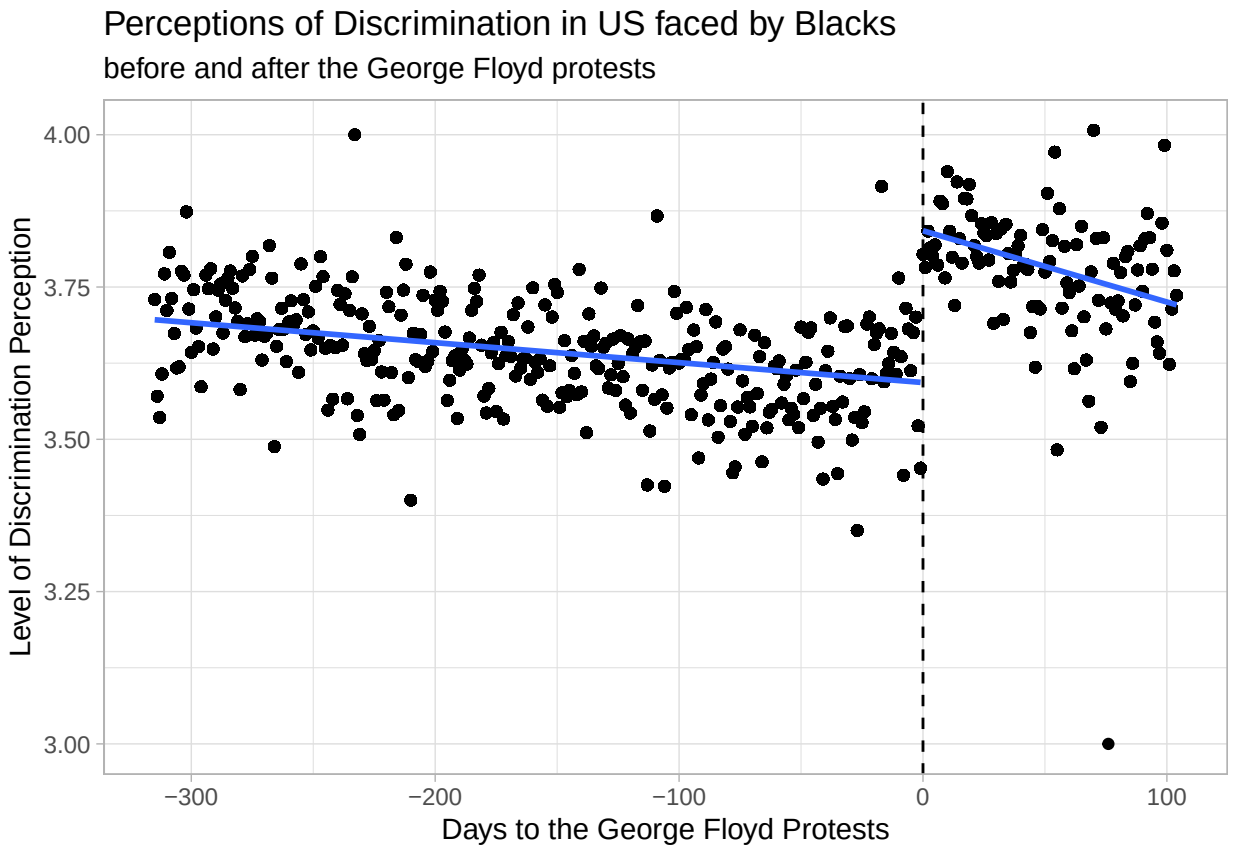
Part c

It might be useful to more clearly illustrate the differences in the trend lines before and after the protests began. **Modify the code below to include a separate line of best fit for before and after the protests began. Does the trend line align with your previous reading of the graph? Remember to add a vertical line demarcating for the onset of treatment.**

```
# group_favorability_the_police
police_favorability_all <- protest_df_bydate %>%
  drop_na(group_favorability_the_police) %>%
  group_by(day_running) %>%
  summarize(avg_group_favorability_the_police = mean(group_favorability_the_police),
            before = before,
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = avg_group_favorability_the_police, group = before)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x", se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  theme_light() +
  labs(
    title = "Favorability Towards The Police",
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Favorability"
  )

police_favorability_all
```




```
# png("police_favorability_all_plot.png", units="in", width=8, height=5, res=300)
# print(police_favorability_all)
# dev.off()

# discrimination_blacks
protest_df_bydate %>%
  drop_na(discrimination_blacks) %>%
  group_by(day_running) %>%
  summarize(avg_discrimination_blacks = mean(discrimination_blacks),
            before = before,
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = avg_discrimination_blacks, group = before)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x", se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  theme_light() +
  labs(
    title = "Perceptions of Discrimination in US faced by Blacks",
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Level of Discrimination Perception"
  )
```



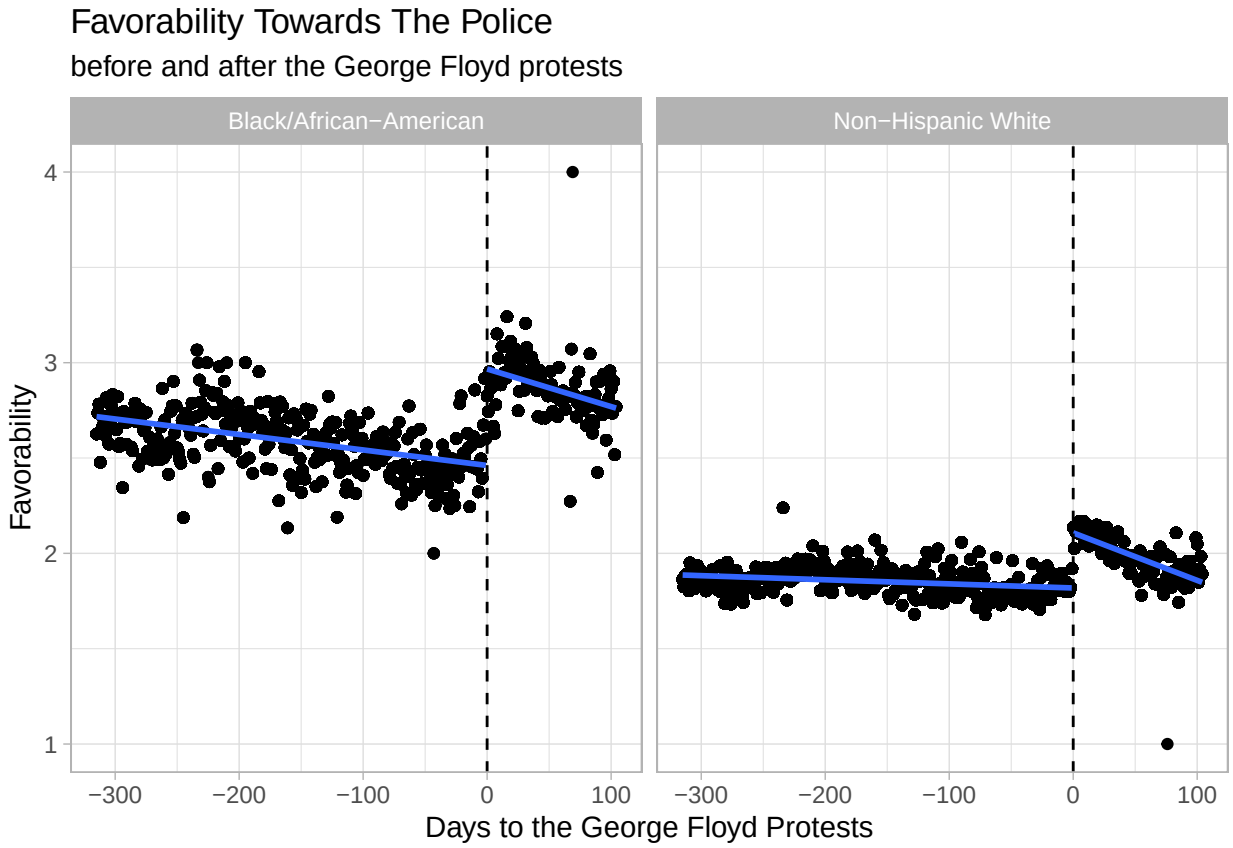
Yes, these trends align with our previous reading of the graph (Higher values represent lower favorability of police and higher levels of discrimination perception).

Question 5

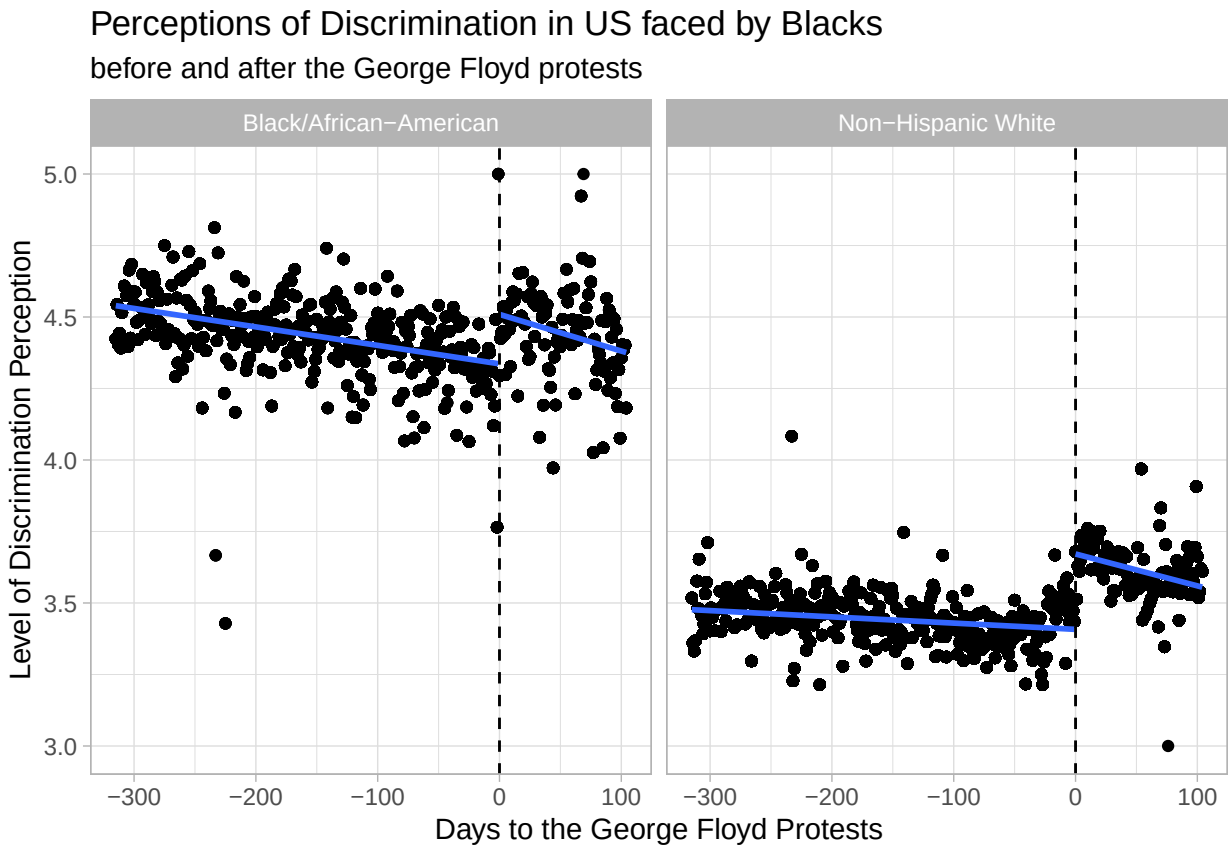
Part a

The attitudes in question are no doubt highly influenced by the respondent's race and ethnicity. How do the graphs from question 4 differ for white and Black respondents. **Subset the data to include only white respondents and recreate the graphs from part c of question 4. Do the same with the data from only Black respondents. How do these differ from each other? Hint: Be careful when subsetting white responses to not also include Hispanic responses.**

```
# group_favorability_the_police
protest_df_byrace_ethnicity %>%
  filter(race_ethnicity_grouped %in% c("Non-Hispanic White", "Black/African-American")) %>%
  drop_na(group_favorability_the_police) %>%
  group_by(day_running, race_ethnicity_grouped) %>%
  summarize(avg_group_favorability_the_police = mean(group_favorability_the_police),
            before = before,
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = avg_group_favorability_the_police, group = before)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x", se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  facet_wrap(~race_ethnicity_grouped) +
  theme_light() +
  labs(
    title = "Favorability Towards The Police",
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Favorability"
  )
```




```
# discrimination_blacks
protest_df_byrace_ethnicity %>%
  filter(race_ethnicity_grouped %in% c("Non-Hispanic White", "Black/African-American")) %>%
  drop_na(discrimination_blacks) %>%
  group_by(day_running, race_ethnicity_grouped) %>%
  summarize(avg_discrimination_blacks = mean(discrimination_blacks),
            before = before,
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = avg_discrimination_blacks, group = before)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x", se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  facet_wrap(~race_ethnicity_grouped) +
  theme_light() +
  labs(
    title = "Perceptions of Discrimination in US faced by Blacks",
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Level of Discrimination Perception"
  )
```



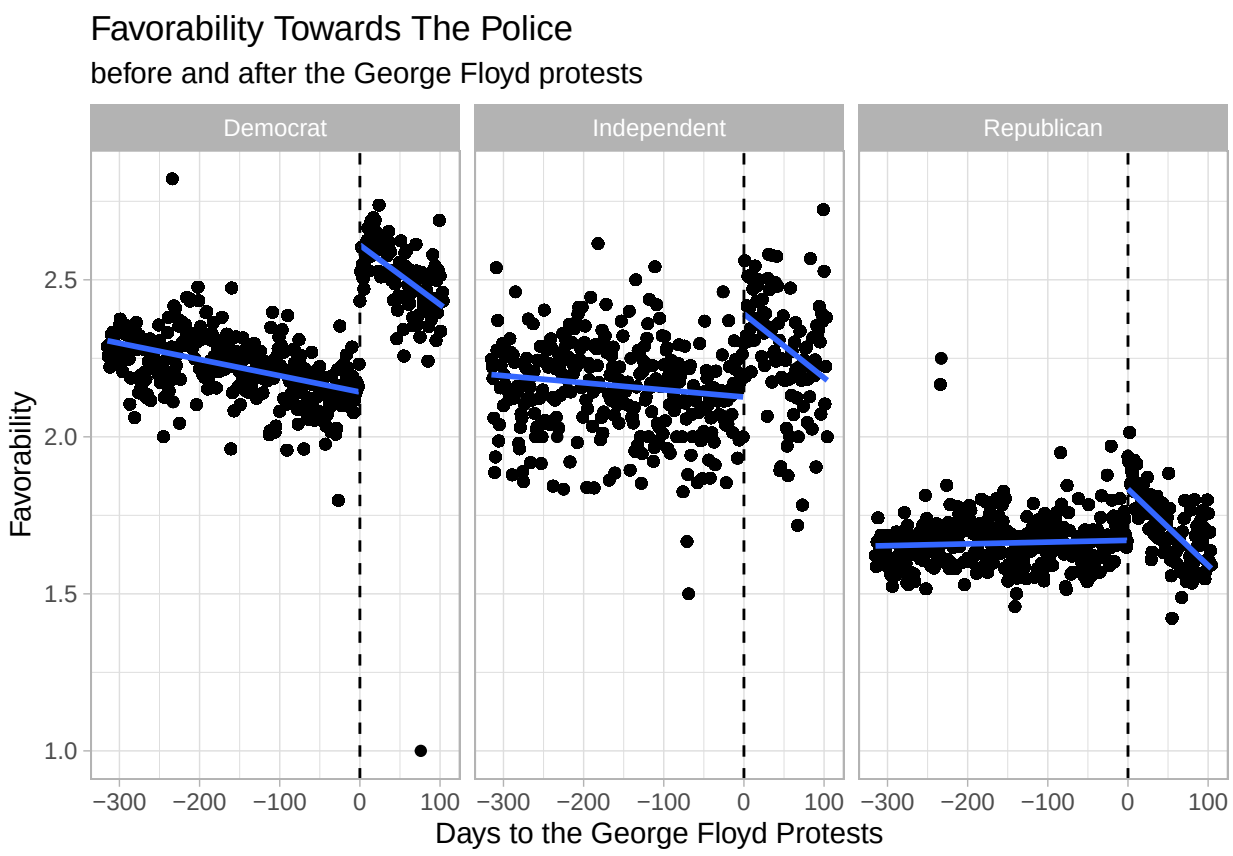
Splitting the trends by race, we can see that on average Black Americans have on average much more unfavorable attitudes towards police and higher levels of discrimination perception than White Americans. After the protests, both groups had a shift towards more unfavorable attitudes towards police and higher levels of discrimination perception. However, White Americans' attitudes towards the police quickly returned to normal levels before the protests after around 100 days, while these attitudes were still much more unfavorable for Black Americans.

Part b

As we have learned partisanship heavily influences how people take in and process new information. **Split the sample into Democrats, Republicans and independents and use them to produce the same graphs as part a (either all in the same figure or separate). Compare both the level and the trends for each party affiliation. What could this imply about how partisanship affects processing?**

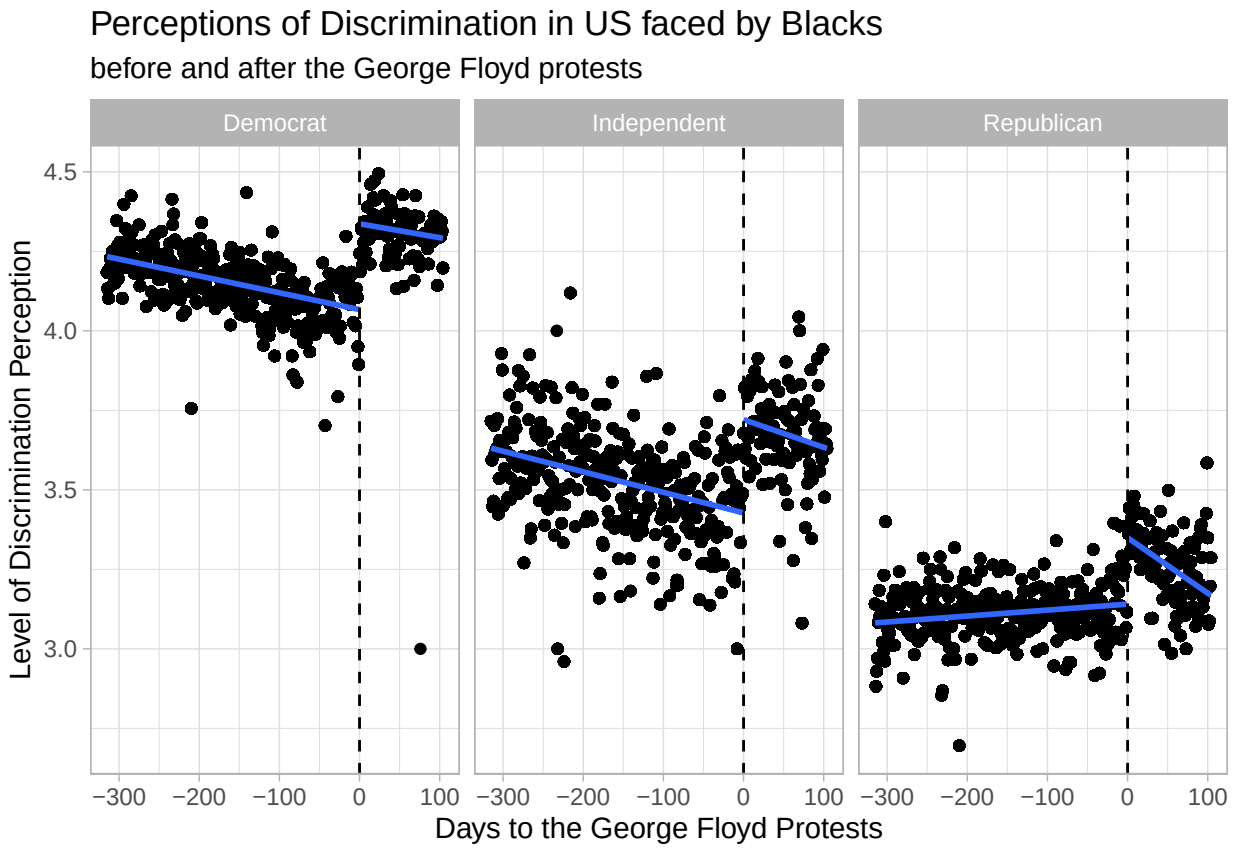
```
# group_favorability_the_police
police_favorability_party <- protest_df_bydate %>%
  mutate(pid3 = case_when(
    pid7 < 4 ~ "Democrat",
    pid7 == 4 ~ "Independent",
    pid7 > 7 ~ NA_character_,
    pid7 > 4 ~ "Republican"
  )) %>%
  drop_na(pid3) %>%
  drop_na(group_favorability_the_police) %>%
  group_by(day_running, pid3) %>%
  summarize(avg_group_favorability_the_police = mean(group_favorability_the_police),
            before = before,
```

```
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = avg_group_favorability_the_police, group = before)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x", se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  facet_wrap(~pid3) +
  theme_light() +
  labs(
    title = "Favorability Towards The Police",
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Favorability"
  )

police_favorability_party
```




```
# png("police_favorability_party_plot.png", units="in", width=8, height=5, res=300)
# print(police_favorability_party)
# dev.off()

# discrimination_blacks
protest_df_bydate %>%
  mutate(pid3 = case_when(
    pid7 < 4 ~ "Democrat",
    pid7 == 4 ~ "Independent",
    pid7 > 7 ~ NA_character_,
    pid7 > 4 ~ "Republican"
```

```
)) %>%
drop_na(pid3) %>%
drop_na(discrimination_blacks) %>%
group_by(day_running, pid3) %>%
summarize(avg_discrimination_blacks = mean(discrimination_blacks),
          before = before,
          .groups = "drop") %>%
ggplot(aes(x = day_running, y = avg_discrimination_blacks, group = before)) +
geom_point() +
geom_smooth(method = "lm", formula = "y ~ x", se = FALSE) +
geom_vline(xintercept = 0, linetype = "dashed") +
facet_wrap(~pid3) +
theme_light() +
labs(
  title = "Perceptions of Discrimination in US faced by Blacks",
  subtitle = "before and after the George Floyd protests",
  x = "Days to the George Floyd Protests",
  y = "Level of Discrimination Perception"
)
```



Splitting the trends by party, we can see that on average Democrats have on average much more unfavorable attitudes towards police and higher levels of discrimination perception than Republicans, with Independents somewhere in between. After the protests, both groups had a shift towards more unfavorable attitudes towards police and higher levels of discrimination perception. However, Republicans' and Independts' attitudes towards the police quickly returned to normal levels before the protests after around 100 days, while these attitudes were still much more unfavorable for Democrats. The same trend holds true for Republicans in relation to the level of discrimination perception, but does not for Independents (which have stayed higher

like the attitudes of Democrats). This tells us that these are still very partisan issues.

Question 6:

Part a

The graphs in questions 4 and 5 indicate that the effects dissipate as time progresses past the onset of the protests. **Explain why that might be the case? What does this indicate about whether or not attitudes towards the police are symbolic or not?**

Attitudes dissipate as time progresses and protests wind down because life appears to return back to normal for most Americans. Humans are also bad at evaluating the past as shown in Healy and Lenz (2014) where voters only respond to the economy of election years compared to the past four years before an election. This indicates that attitudes towards the police are not symbolic and do shift. In comparison, symbolic variables such as party affiliation and ideology remained the same before and after the protests in the survey.

Part b

One way to look at the effect decay is to bin the post-protest data and compare averages. **Split the post-protest data into however many groups you choose and compare the period directly after the protest with the latest period in the data. What are the differences in means for the outcomes?**

```
before_df <- protest_df_bydate %>%
  drop_na(group_favorability_the_police) %>%
  group_by(day_running) %>%
  summarize(avg_group_favorability_the_police = mean(group_favorability_the_police),
            before = before,
            .groups = "drop") %>%
  filter(before == 0)

after_df <- protest_df_bydate %>%
  drop_na(group_favorability_the_police) %>%
  group_by(day_running) %>%
  summarize(avg_group_favorability_the_police = mean(group_favorability_the_police),
            before = before,
            .groups = "drop") %>%
  filter(before == 1)

lm(formula = avg_group_favorability_the_police ~ day_running, data = before_df)
```

```
##
## Call:
## lm(formula = avg_group_favorability_the_police ~ day_running,
##     data = before_df)
##
## Coefficients:
## (Intercept)  day_running
##    2.275102    -0.002279
```

```
lm(formula = avg_group_favorability_the_police ~ day_running, data = after_df)
```

```
##
## Call:
## lm(formula = avg_group_favorability_the_police ~ day_running,
##     data = after_df)
##
## Coefficients:
## (Intercept)  day_running
```

```
##   1.9556633   -0.0002967
```

$$
\begin{aligned}
2.275102 - 0.002279x &= 1.9556633 - 0.0002967x \\
0.3194387 &= 0.0019823x \\
x &= 161.145487565
\end{aligned}
$$

In section, I calculated the two lm trend lines for attitudes towards the police before and after the protests. Based on these two fits, I then calculated on which day they would intercept, which is around 161 days after the protests. This makes sense to me, as the average attitudes had already begun to bounce back to the pre-protest levels 100 days after the protests.

Question 7

Part a

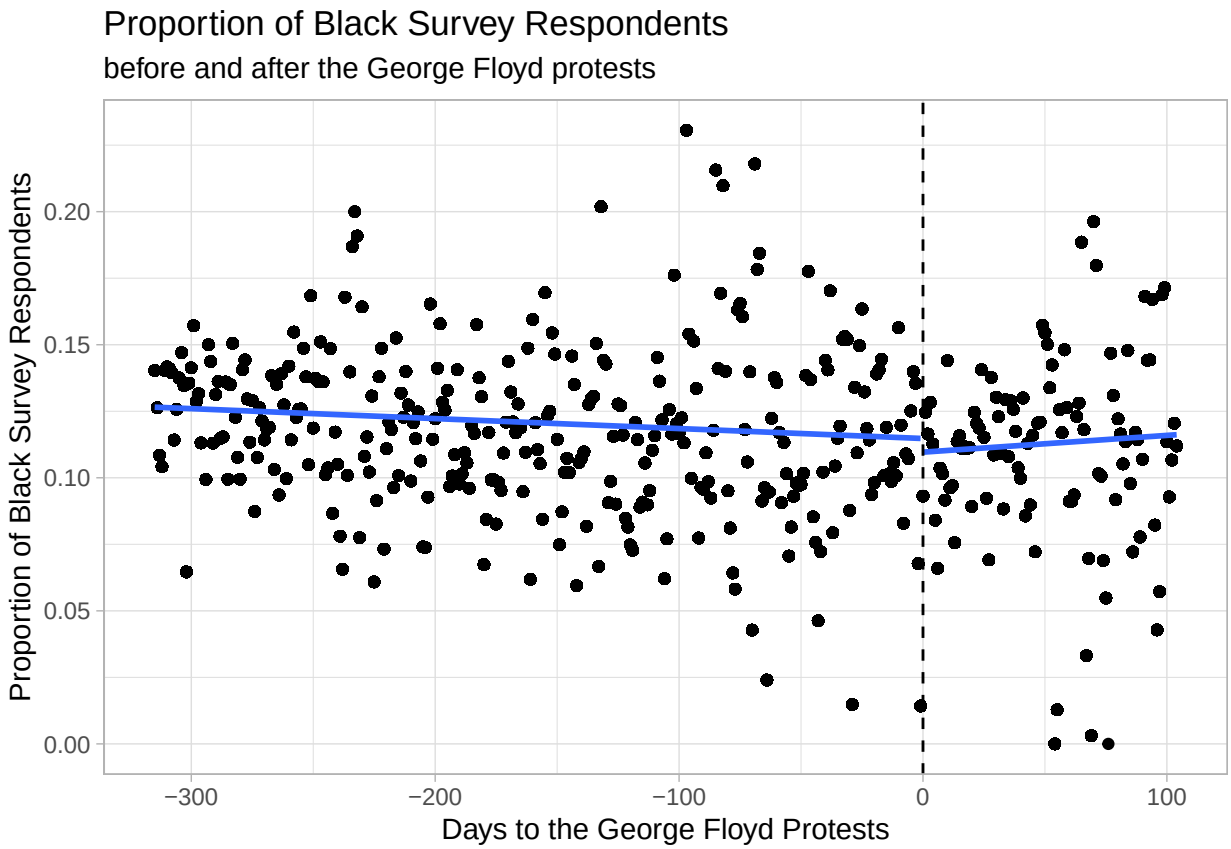
What are some reasons we might be unconvinced by the comparison of aggregate survey results from a time before and after an event? Do you think they apply here?

If the sample was different and we were comparing surveys conducted with different methods (phone vs internet), the I would have many concerns including a different group of representation coming from coverage error. However, since this data is from the Nationscape survey (NS) conducted by the Democracy Fund and UCLA, I know that the researchers have taken considerable time to ensure that coverage error is minimized.

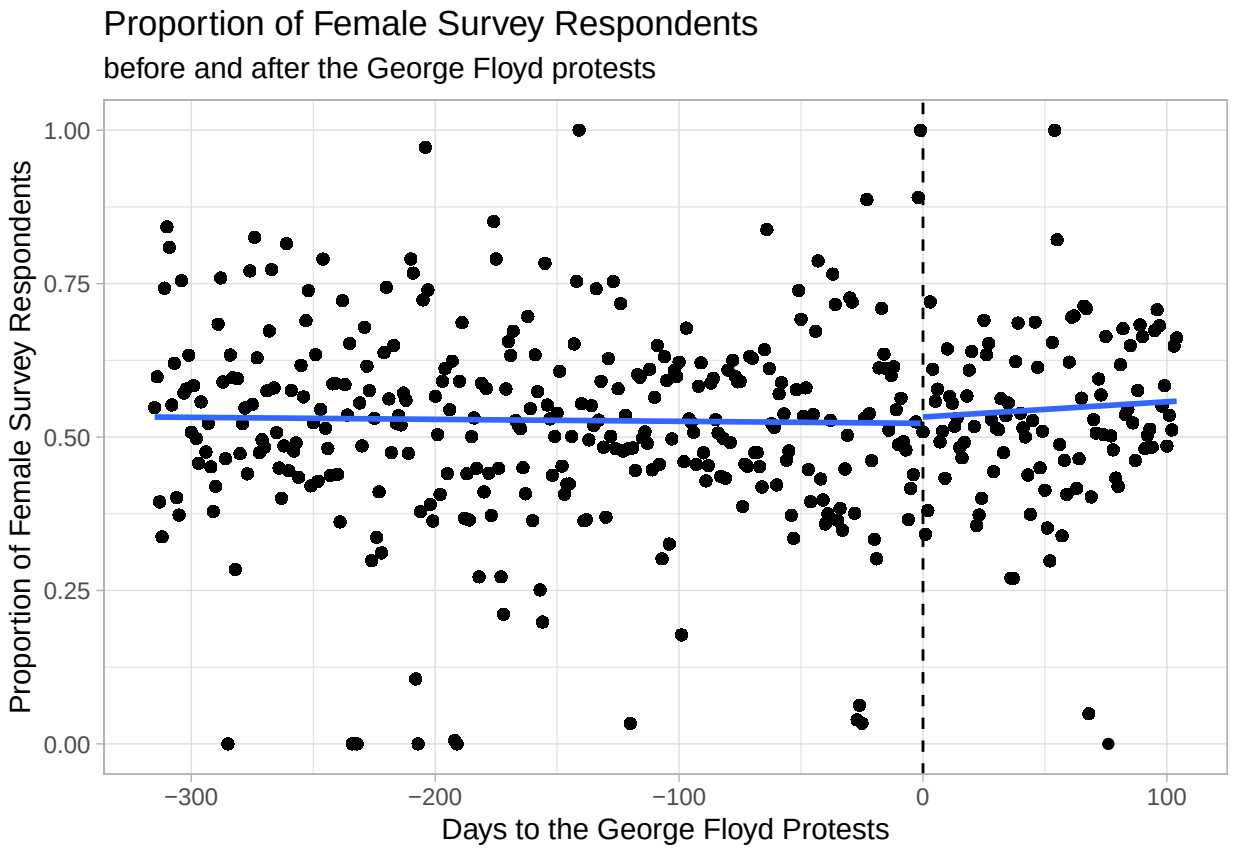
Part b

There is often a problem in conducting surveys of non-response bias. That is, the people who answer surveys may differ from the people who do not answer surveys and the differences may vary over time. This is especially damaging to inference when non-response is correlated with the outcomes being measured. For example after a series of damaging headlines supporters of a politician may be less willing to answer phone surveys about that politician. As a result we would potentially observe an exaggeration of the negative effects of the scandal on a politician's polled approval rating. Test whether this is the case in the Reny and Newman data. **Test whether there is balance between the respondents before and after the onset of the protests along two demographic traits that you would expect to correlate with the measured responses to the outcome variables.**

```
# Proportion Black Respondents
protest_df_bydate %>%
  mutate(black = (race_ethnicity == 2)) %>%
  group_by(day_running) %>%
  summarize(prop_black = sum(black) / n(),
            before = before,
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = prop_black, group = before)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x", se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  theme_light() +
  labs(
    title = "Proportion of Black Survey Respondents",
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Proportion of Black Survey Respondents"
  )
```




```
# Proportion Female Respondents
protest_df_bydate %>%
  group_by(day_running) %>%
  summarize(prop_female = sum(female) / n(),
            before = before,
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = prop_female, group = before)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x", se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  theme_light() +
  labs(
    title = "Proportion of Female Survey Respondents",
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Proportion of Female Survey Respondents"
  )
```



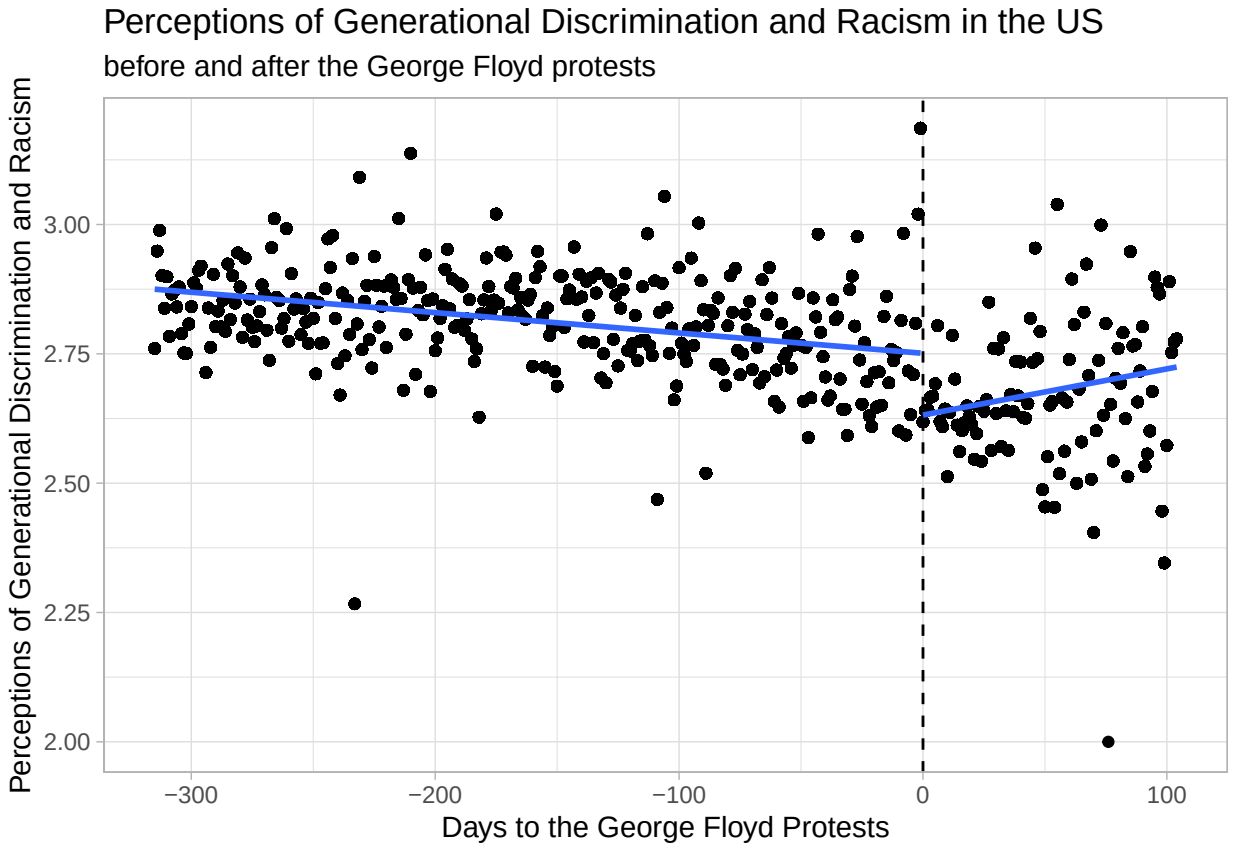
In this case, I tested the proportion of respondents that were Black and Female before and after the protests. We can see that non-response bias was not increased after the protests, which might have resulted in the proportion of these respondents to decrease after the protests. Instead, we see little change and even a small increase.

Part c

Racial resentment is often considered a symbolic attitude in strength and consistency. Examine the before and after levels of racial resentment as measured by the question from the racial resentment scale about the impact of generations of slavery and discrimination (racial_attitudes_generations). **Graph the average racial_attitudes_generations (remember the direction of how it is coded!) by day like other outcome variables. Does it behave like the other outcome variables? Does the data support that racial attitudes are symbolic attitudes?**

```
# racial_attitudes_generations
protest_df_bydate %>%
  drop_na(racial_attitudes_generations) %>%
  group_by(day_running) %>%
  summarize(avg_racial_attitudes_generations = mean(racial_attitudes_generations),
            before = before,
            .groups = "drop") %>%
  ggplot(aes(x = day_running, y = avg_racial_attitudes_generations, group = before)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x", se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  theme_light() +
  labs(
    title = "Perceptions of Generational Discrimination and Racism in the US",
```

```
    subtitle = "before and after the George Floyd protests",
    x = "Days to the George Floyd Protests",
    y = "Perceptions of Generational Discrimination and Racism"
  )
```



The trends in this variable are different than the others we've looked at so far. In the case of generational slavery and racism, we see that more people appear to disagree with the impact that generational slavery and racism has post-protests. However, we can also see that the responses tended to be much more variable after the protests began and much more heteroskedastic. In fact, the general trend seems to have stayed relatively stable on average besides the increase in variability. This data might not support the fact that racial attitudes are symbolic attitudes since the attitudes seemed to have changed. However, the variability also might show that racial attitudes are not symbolic and have the potential to change.

Question 8: Data Science Question

Part a

Run an initial regression examining the relationship between favorability towards the police, party, and treatment. **Run a regression examining party and the onset of the protests' effect on favorability towards the police. Interpret the results**

```
fit_1 <- lm(group_favorability_the_police ~ pid7 + before, data = protest_df_bydate)

stargazer(fit_1, header = FALSE,
          dep.var.labels = c("Police Favorability"),
          covariate.labels = c("Party ID", "Before Protests Onset", "Constant"),
          title = "Police Favorability as a Function of Party ID and Protests")
```

Table 2: Police Favorability as a Function of Party ID and Protests

	Dependent variable:
	Police Favorability
Party ID	-0.127^{***}
	(0.001)
Before Protests Onset	-0.169^{***}
	(0.004)
Constant	2.646^{***}
	(0.004)
Observations	326,797
R^2	0.088
Adjusted R^2	0.088
Residual Std. Error	0.966 (df = 326794)
F Statistic	$15{,}807.950^{***}$ (df = 2; 326794)
Note:	*p<0.1; **p<0.05; ***p<0.01

From the model, we can see that a one point increase in party ID (towards Republican) is associated with a decrease on average of 0.127 on the police favorability scale (lower is more favorable). Attitudes, on average, across all respondents were 0.169 more favorable towards the police before the onset of the protests. Both of these results are statistically significant at the 95% level with p-values below 0.01.

Part b

The above functional form probably does not accurately model the relationship of all the relevant covariates in the dataset. What functional form would you recommend using and why? What covariates would you add? Is there need for an interaction term? **Run a regression of your specificiation and interpret the results. Justify your choices in modeling.**

```
logit_fit <- polr(as.factor(group_favorability_the_police) ~ pid7 + before,
                  method = "logistic", data = protest_df_bydate)

probit_fit <- polr(as.factor(group_favorability_the_police) ~ pid7 + before,
                  method = "probit", data = protest_df_bydate)

stargazer(fit_1, logit_fit, probit_fit, header = FALSE,
```

```
dep.var.labels = c("Police Favorability", "Police Favorability"),
covariate.labels = c("Party ID", "Before Protests Onset", "Constant"),
title = "Police Favorability as a Function of Party ID and Protests")
```

Table 3: Police Favorability as a Function of Party ID and Protests

	Dependent variable:		
	Police Favorability	Police Favorability	
	OLS	*ordered logistic*	*ordered probit*
	(1)	(2)	(3)
Party ID	−0.127***	−0.259***	−0.150***
	(0.001)	(0.001)	(0.001)
Before Protests Onset	−0.169***	−0.298***	−0.178***
	(0.004)	(0.007)	(0.004)
Constant	2.646***		
	(0.004)		
Observations	326,797	326,797	326,797
R^2	0.088		
Adjusted R^2	0.088		
Residual Std. Error	0.966 (df = 326794)		
F Statistic	15,807.950*** (df = 2; 326794)		
Note:			*p<0.1; **p<0.05; ***p<0.01

Since OLS regressions are not the best for ordinal data analysis, I instead chose to compare the results of a ordered logit and probit model against the OLS results. Here, we can see that the coefficients from the logit and probit models still show the same genreal trend. However, the coefficient effects for the logit model are about twice as large as the OLS model and the effects for the probit model are just slightly larger. This difference is due to the logit model using a cumulative standard logistic distribution and the probit model using a cumulative standard normal distribution. If I were to expand upon these model, I might perhaps look at implementing a heteroskedastic probit model to account for non-constant error variances (the increase in variability of responses) that might alter the magnitude of effects.

Part c

Linear models are not well suited for bounded ordinal responses. Instead ordinal logit or probit models are frequently employed in order to capture a) that the outcomes are restricted to a scale (in the case of police unfavorability 1-4) and b) that the differences between different rungs on the scale are not necessarily equivalent (going from very unfavorable to somewhat unfavorable is not necessarily the same difference as going from somewhat unfavorable to somewhat favorable). **Using the code below from the MASS package run an ordinal probit model using the same model as part b. How do the coefficients differ from part b?**

See part b.