Quick EDA

Victor Møller Poulsen

Load packages

load data & mutate new columns

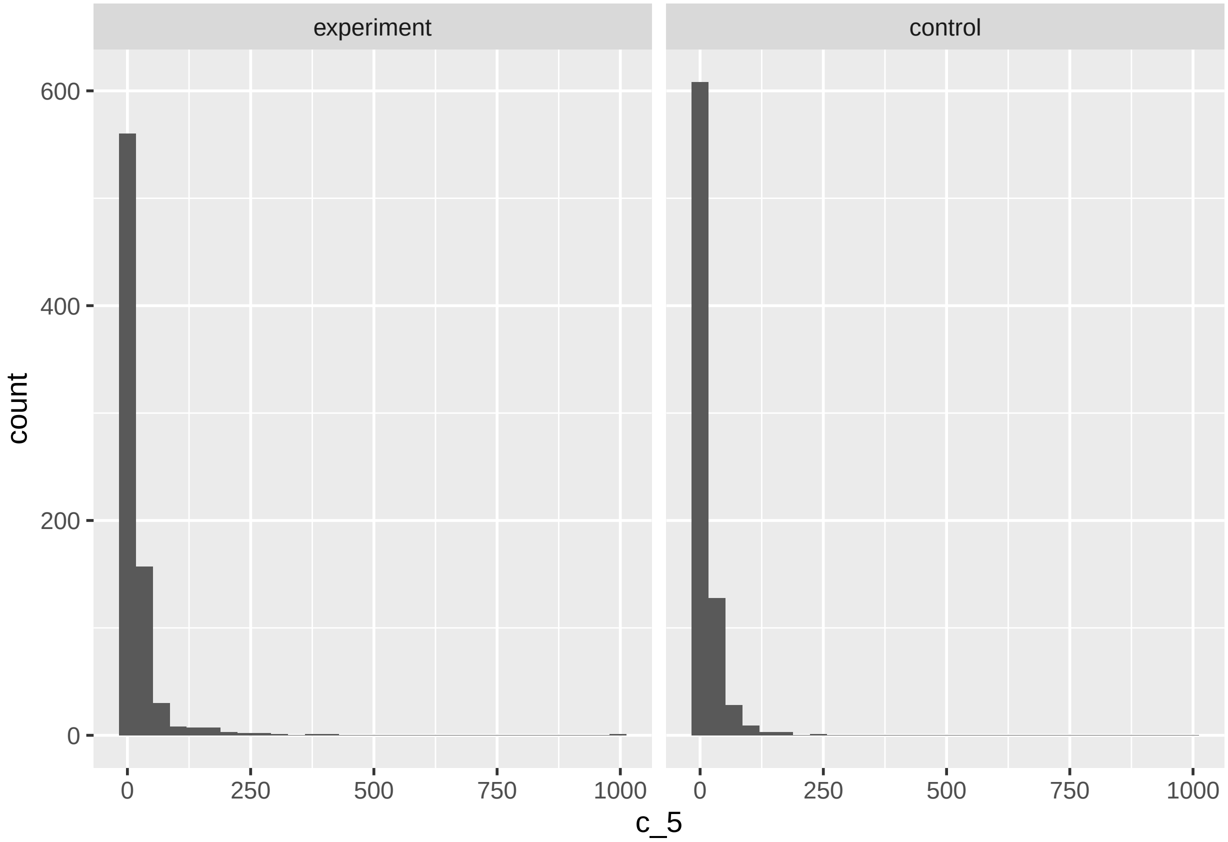
```
# mutate new columns
d <- read_csv("/work/50114/MAG/data/modeling/psych_replication_matched.csv") %>%
  mutate(log_teamsize = log(n_authors),
         condition_coded = ifelse(condition == "experiment", 1, 0),
         condition_fct = as_factor(condition),
         teamsize_scaled = (n_authors-min(n_authors))/(max(n_authors)-min(n_authors)),
         days_after_2010_scaled = days_after_2010/max(days_after_2010),
         teamsize_log = log(n_authors),
         id_match = as_factor(match_group),
         id_fct = as_factor(PaperId)) %>% # because min = 0
  glimpse()
```

```
## Rows: 1560 Columns: 6

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): condition
## dbl (5): match_group, n_authors, PaperId, days_after_2010, c_5

##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.

## Rows: 1,560
## Columns: 14
## $ match_group            <dbl> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8,~
## $ condition              <chr> "experiment", "control", "control", "experiment~
## $ n_authors              <dbl> 3, 3, 1, 1, 4, 4, 5, 5, 2, 2, 2, 2, 3, 3, 5, 5,~
## $ PaperId                <dbl> 2330249536, 2003350634, 2385753682, 2395494269,~
## $ days_after_2010        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ c_5                    <dbl> 10, 0, 0, 0, 310, 0, 2, 17, 0, 13, 2, 13, 0, 0,~
## $ log_teamsize           <dbl> 1.0986123, 1.0986123, 0.0000000, 0.0000000, 1.3~
## $ condition_coded        <dbl> 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1,~
## $ condition_fct          <fct> experiment, control, control, experiment, exper~
## $ teamsize_scaled        <dbl> 0.03333333, 0.03333333, 0.00000000, 0.00000000,~
## $ days_after_2010_scaled <dbl> 0.000000000, 0.000000000, 0.000000000, 0.000000~
## $ teamsize_log           <dbl> 1.0986123, 1.0986123, 0.0000000, 0.0000000, 1.3~
## $ id_match               <fct> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8,~
## $ id_fct                 <fct> 2330249536, 2003350634, 2385753682, 2395494269,~
```

Check distributions

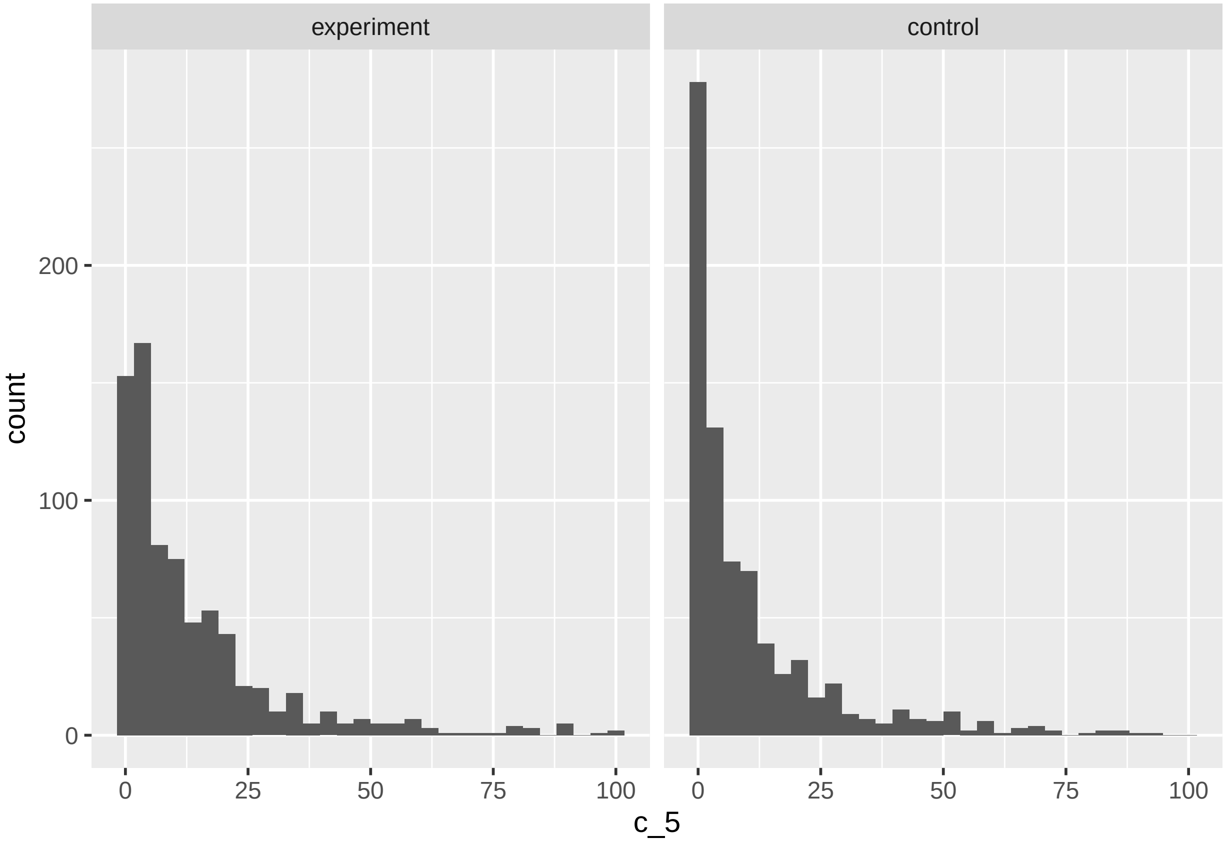
```
d %>% ggplot(aes(x = c_5)) +
  geom_histogram() +
  facet_wrap(vars(condition_fct))
```

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



```
d %>% filter(c_5 <= 100) %>%
  ggplot(aes(x = c_5)) +
  geom_histogram() +
  facet_wrap(vars(condition_fct))
```

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



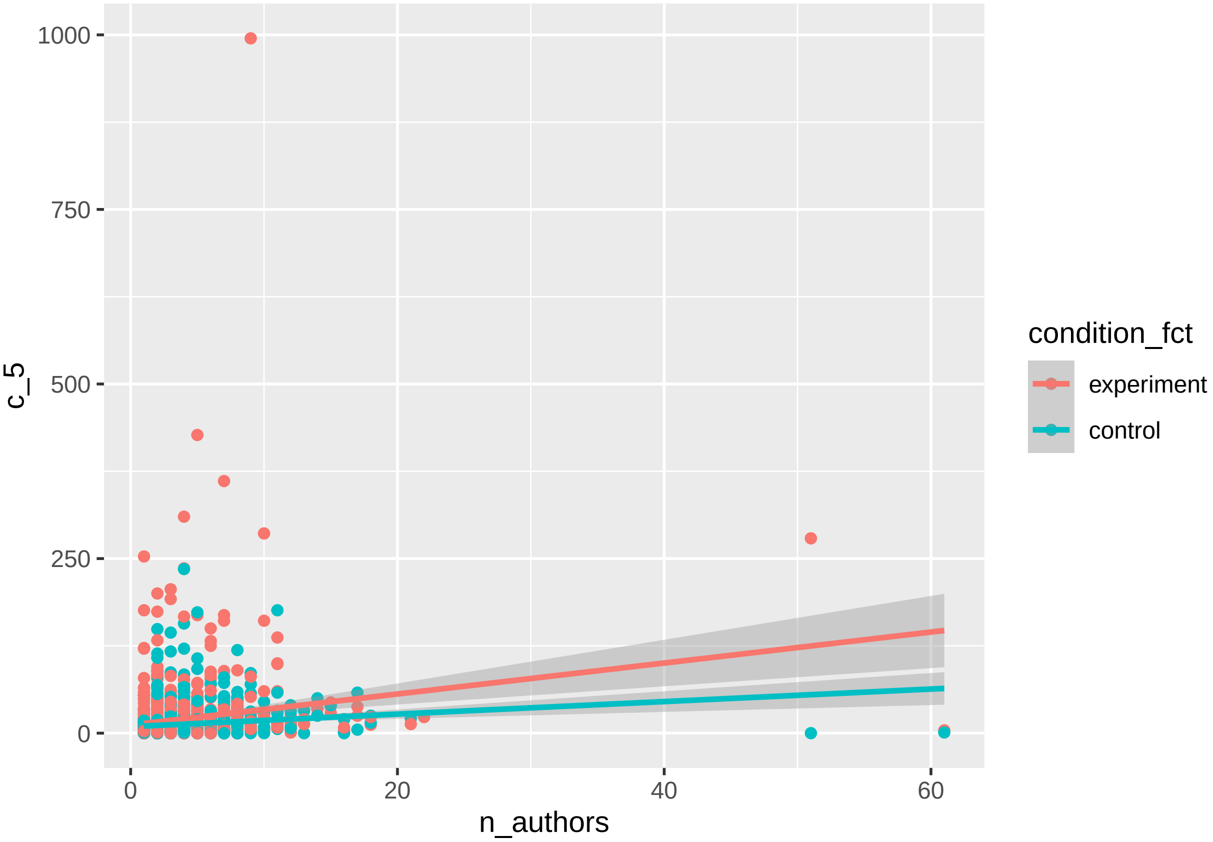
Check teamsize interaction with condition

raw

strong effect of teamsize, interaction probably driven by few data-points

```
d %>% ggplot(aes(x = n_authors, y = c_5, color = condition_fct)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

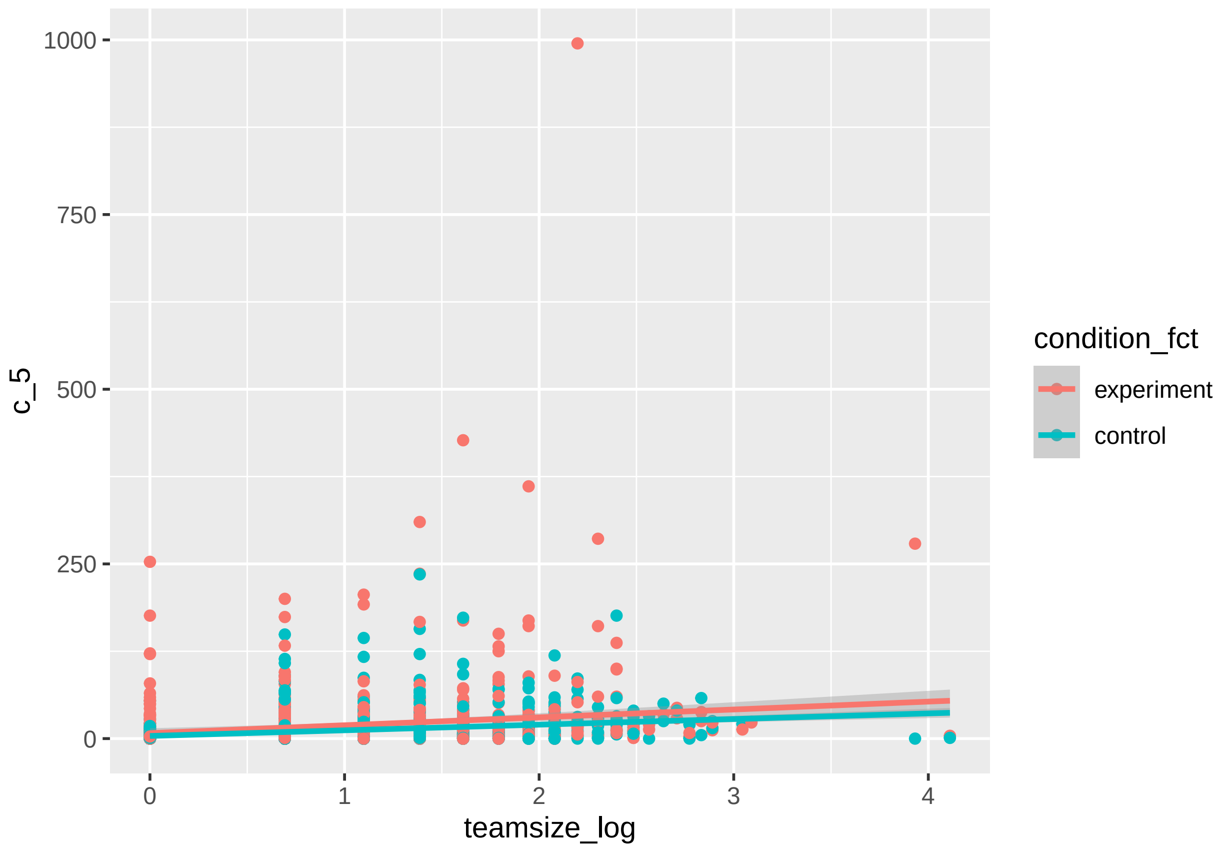


log

not sure whether this is more appropriate

```
d %>% ggplot(aes(x = teamsize_log, y = c_5, color = condition_fct)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

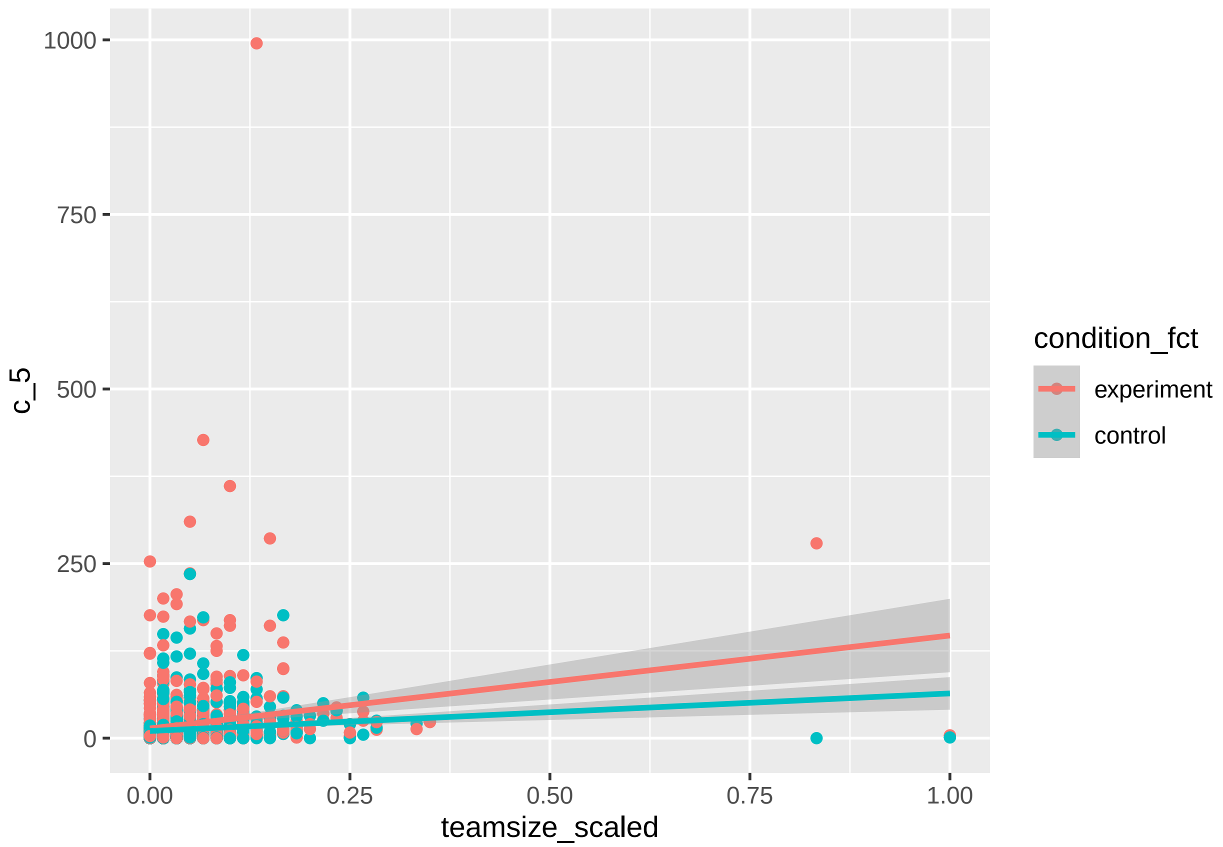


0-1 scaling

Same as original, just might be better for sampling.

```
d %>% ggplot(aes(x = teamsize_scaled, y = c_5, color = condition_fct)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

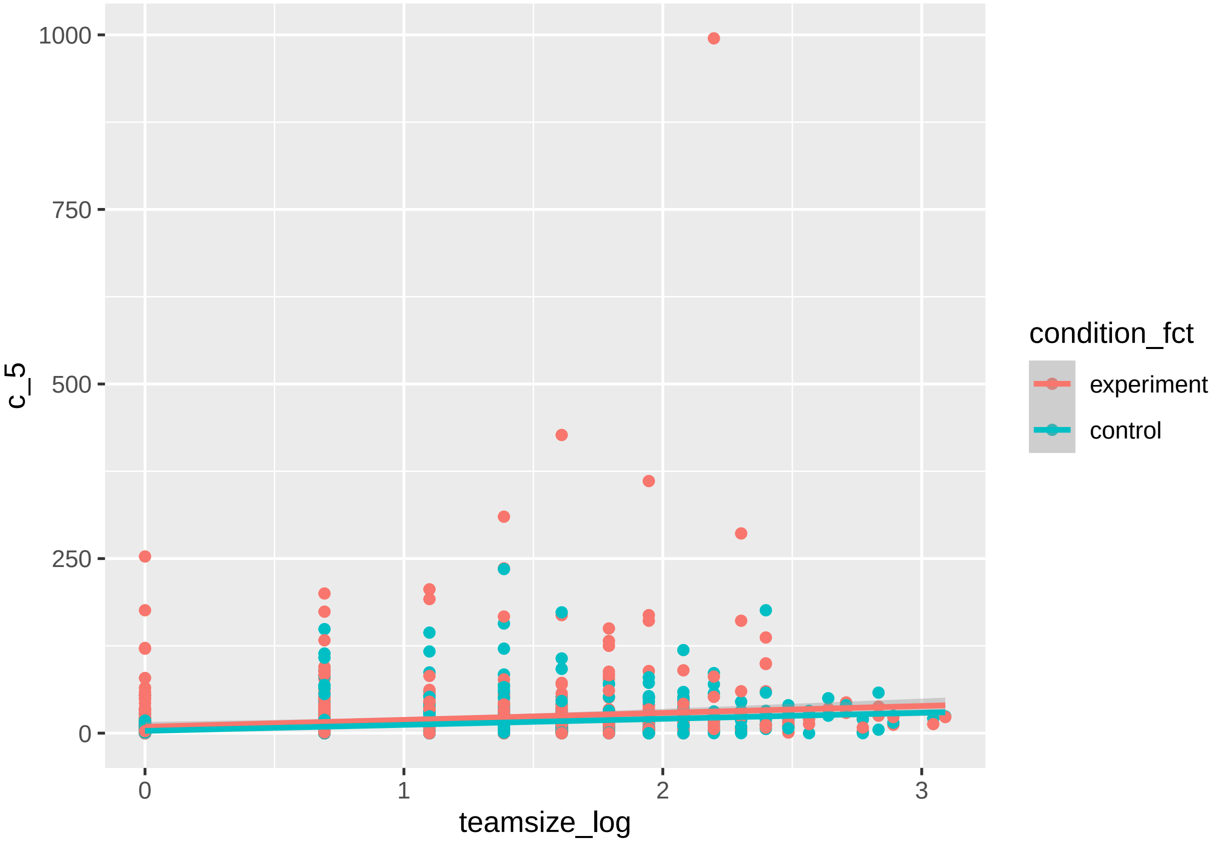


log without the two highest points

Clearly, no interaction effect of teamsize and condition. There is just a main effect of teamsize & a (very slight) main effect of condition.

```
d %>% filter(teamsize_log < 3.5) %>%
  ggplot(aes(x = teamsize_log, y = c_5, color = condition_fct)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

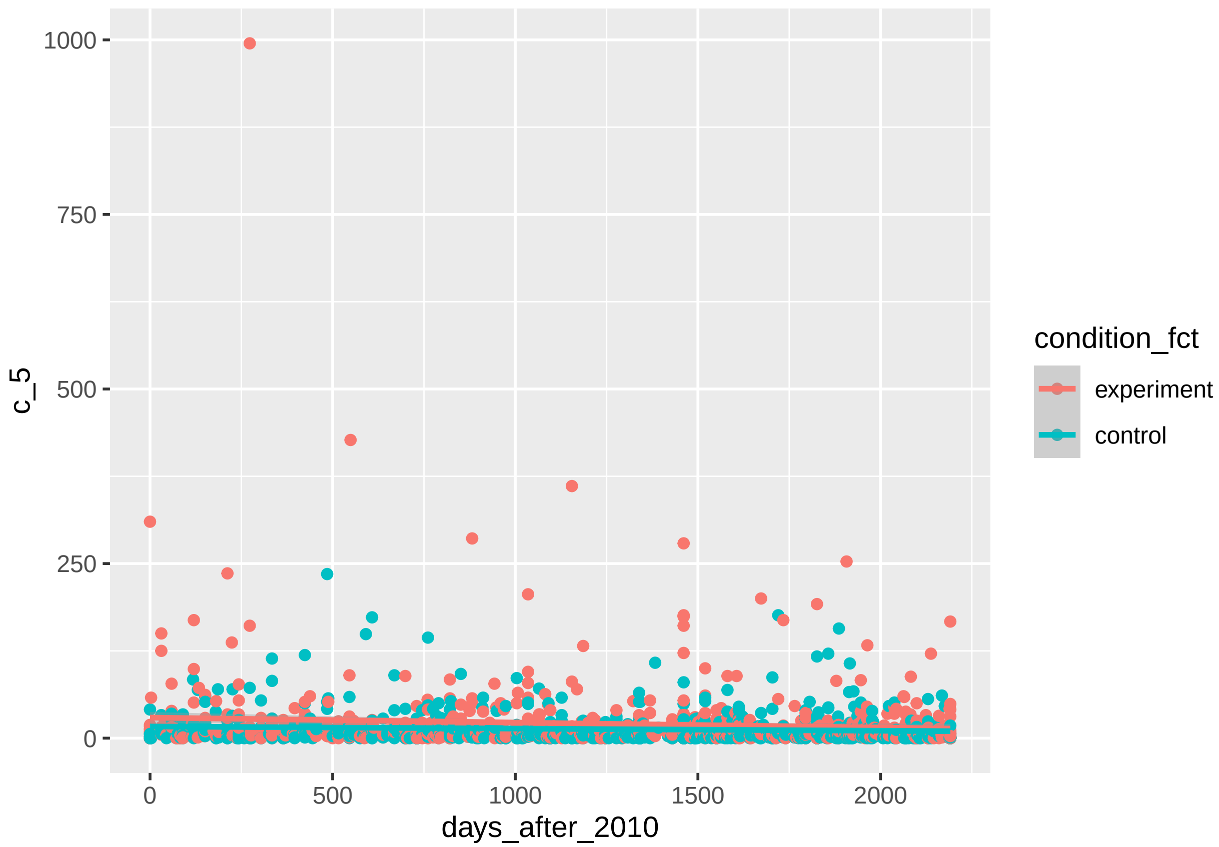


check days interaction

Also clealy no interaction effect. Slightly decreasing trend, not interesting. Probably just want to control for this.

```
d %>%
  ggplot(aes(x = days_after_2010, y = c_5, color = condition_fct)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```


c_5
1000
750
500
250
0
0
500
1000
1500
2000
days_after_2010
condition_fct
experiment
control