cvd_region_confounder

Zhongyi Guo

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
##
## The following object is masked from 'package:stats':
##
##     filter
```

Exploratory Data Analysis

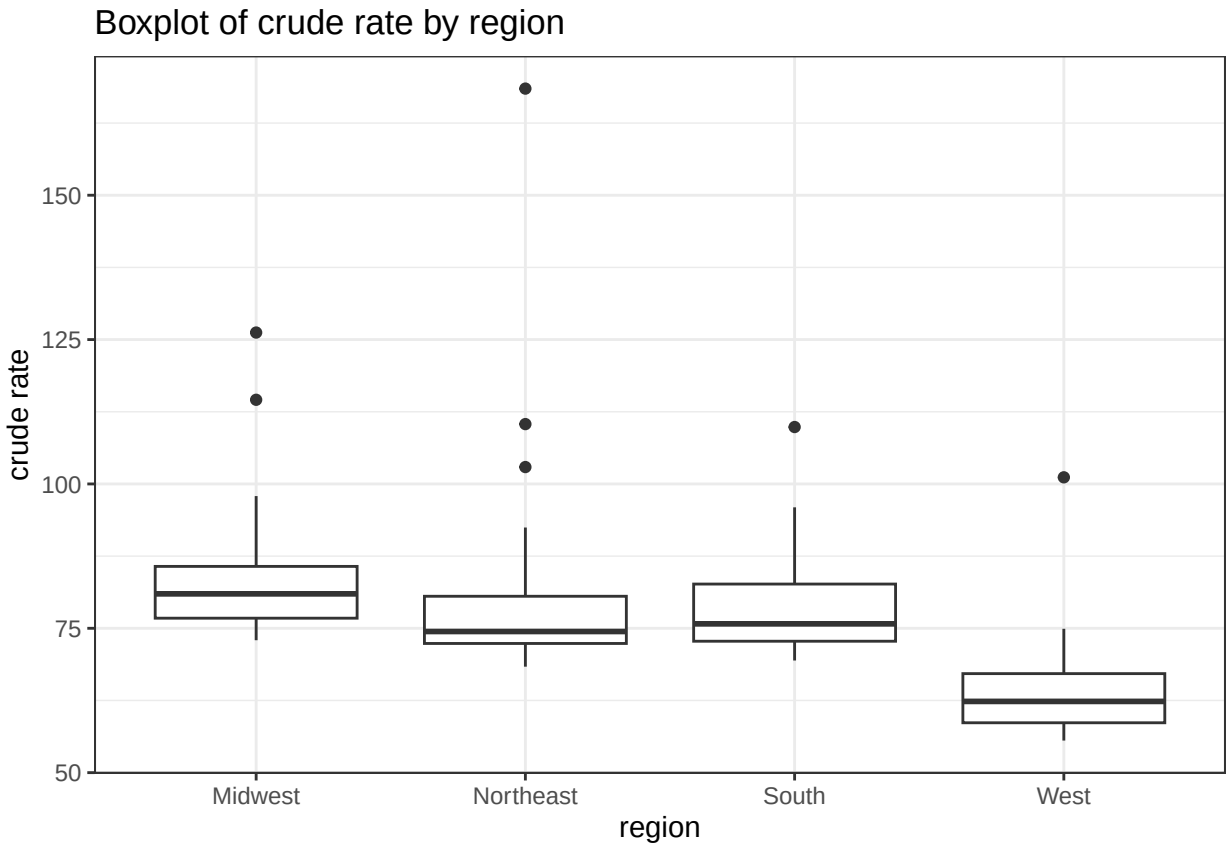
```
# load region data
region <- read.csv("../data/cvd_region_crude_rate.csv")

# create a dummy variable `covid` to indicate if this row of data is
# before COVID or not. If before COVID happened (Pre-COVID), covid = 0; if after
# COVID happened, covid = 1
before_covid_index <- append(grep("2018", region$month), grep("2019",
                                                               region$month))
region$covid <- 1
region$covid[before_covid_index] <- 0
```

```
# report summary statistics of `crude_rate`
region %>%
  group_by(region) %>%
  summarise(
    count = n(),
    mean = mean(crude_rate, na.rm = TRUE),
    sd = sd(crude_rate, na.rm = TRUE),
    median = median(crude_rate, na.rm = TRUE),
    IQR = IQR(crude_rate, na.rm = TRUE)
```

```
    )
```

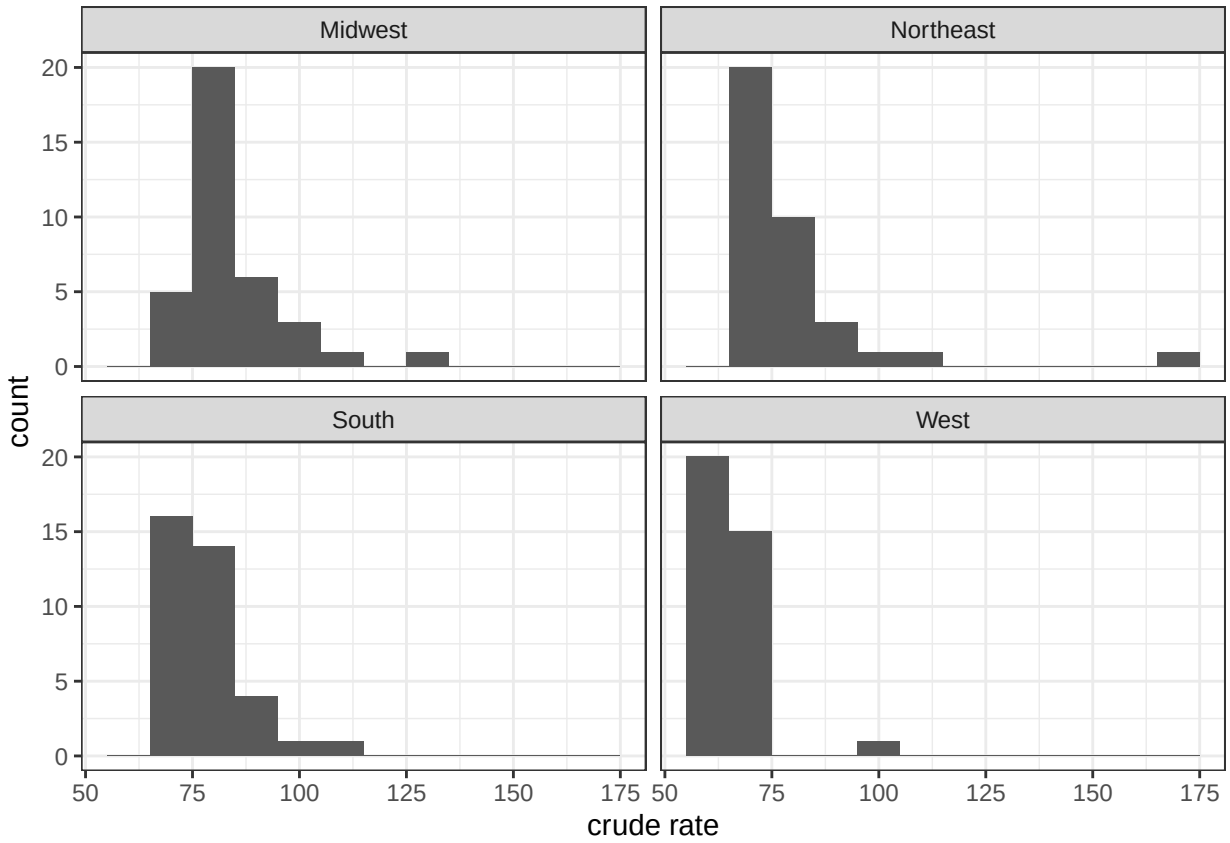
```
## # A tibble: 4 x 6
##   region    count  mean    sd median   IQR
##   <chr>     <int> <dbl> <dbl>  <dbl> <dbl>
## 1 Midwest      36  84.0 11.1    81.0  8.97
## 2 Northeast    36  79.9 17.7    74.4  8.18
## 3 South        36  79.2  8.93   75.8  9.90
## 4 West         36  64.4  8.24   62.3  8.51
```

```
# create a boxplot for crude rates of each region
ggplot(region, aes(x = region, y = crude_rate)) + geom_boxplot() +
  labs(x = "region", y = "crude rate",
       title = "Boxplot of crude rate by region") + theme_bw()
```



The boxplot indicates some outliers in all regions. We will save them for now and remove them later.

```
# create histograms for each region, wrapped in facet
region %>%
  group_by(region) %>%
  ggplot(aes(x=crude_rate)) + geom_histogram(binwidth = 10) +
  facet_wrap(~ region, nrow = 2, ncol = 2) +
  labs(x = "crude rate") +
  theme_bw()
```



Normality test

```
# report p-values of Shapiro test for each region
region %>%
  group_by(region) %>%
  summarize(p_value = shapiro.test(crude_rate)$p.value)
```

```
## # A tibble: 4 x 2
##   region          p_value
##   <chr>             <dbl>
## 1 Midwest   0.00000371
## 2 Northeast 0.00000000229
## 3 South     0.000238
## 4 West      0.00000226
```

The p-value of the Shapiro normality test for each region are all smaller than 0.05, which indicates that the crude rate of all regions is not normally distributed. Thus, for the next step, we will first detect and remove outliers of each region.

Outlier Removal

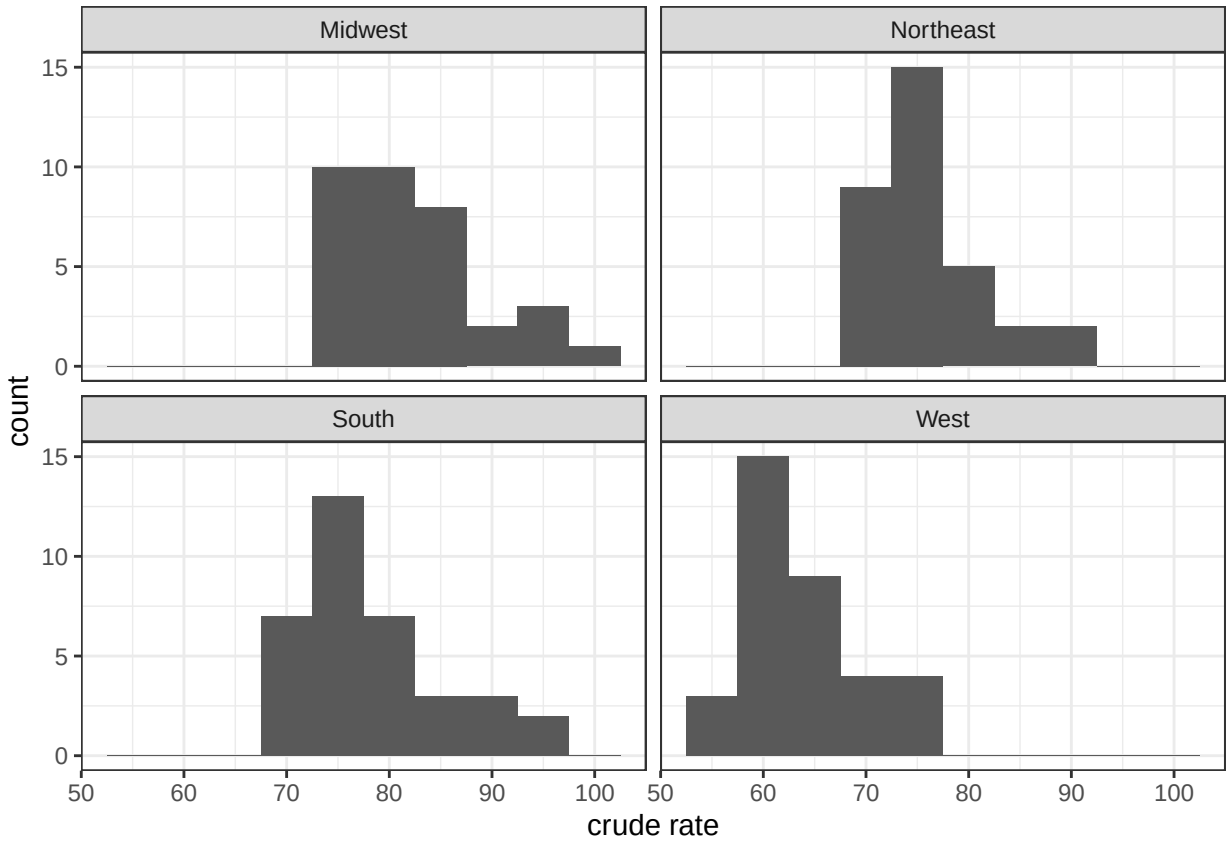
```
# find outliers of each region
region_outliers <- region %>%
  group_by(region) %>%
  identify_outliers(crude_rate)
region_outliers
```

```
## # A tibble: 7 x 8
##   region    month   crude_rate  death population covid is.outlier is.extreme
##   <chr>     <chr>        <dbl>  <int>      <int> <dbl> <lgl>      <lgl>
## 1 Midwest   2020/11       115.  78273   68316744     1 TRUE       TRUE
## 2 Midwest   2020/12       126.  86227   68316744     1 TRUE       TRUE
## 3 Northeast 2020/04       168.  94088   55849869     1 TRUE       TRUE
## 4 Northeast 2020/05       103.  57482   55849869     1 TRUE       FALSE
## 5 Northeast 2020/12       110.  61639   55849869     1 TRUE       TRUE
## 6 South     2020/12       110. 139145  126662754     1 TRUE       FALSE
## 7 West      2020/12       101.  79557   78654756     1 TRUE       TRUE
```

```
# remove outliers for each region
region <- region %>%
  anti_join(region_outliers)
```

```
## Joining with `by = join_by(region, month, crude_rate, death, population,
## covid)`
```

```
# create histograms for each region, wrapped in facet
region %>%
  group_by(region) %>%
  ggplot(aes(x=crude_rate)) + geom_histogram(binwidth = 5) +
  facet_wrap(~ region, nrow = 2, ncol = 2) +
  labs(x = "crude rate") +
  theme_bw()
```




```
# perform normality test again after outlier removal
region %>%
```

```
  group_by(region) %>%
  shapiro_test(crude_rate)
```

```
## # A tibble: 4 x 4
##   region    variable   statistic       p
##   <chr>     <chr>          <dbl>   <dbl>
## 1 Midwest   crude_rate     0.918 0.0146
## 2 Northeast crude_rate     0.900 0.00520
## 3 South     crude_rate     0.892 0.00249
## 4 West      crude_rate     0.925 0.0193
```

The result shows that the crude rate of each region is not normally distributed, after outlier removal. We will then perform a Wilcox test for rank sum.

```
pairwise.wilcox.test(region$crude_rate, region$region, p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum exact test
##
## data:  region$crude_rate and region$region
##
##           Midwest Northeast South
## Northeast 1.7e-05 -         -
## South     0.011   0.142     -
## West      < 2e-16 6.5e-13   1.4e-14
##
## P value adjustment method: none
```

We can conclude Midwest and Northeast, Midwest and South, Midwest and West, Northeast and West, and South and West have strong evidence to suggest that these groups have significantly different distributions because of smaller than 0.05 p-values. There is insufficient evidence to conclude that there is a significant difference between Northeast and South.

This observation can be further explored in discussion with regards to healthcare systems, weather, economic levels, etc., of each region.

Now, the dataset `region` is clean. We will save it as a new file named `cvd_region_crude_rate.csv` and reimport it.

```
write.csv(region, file = "../data/cvd_region_crude_rate_clean.csv",
          row.names = FALSE)
region <- read.csv("../data/cvd_region_crude_rate_clean.csv")
```

ANOVA - Are crude rate means different among regions?

```
anova_result <- region %>%
  aov(crude_rate ~ region, data = .)
summary(anova_result)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## region        3   6750  2249.9   56.36 <2e-16 ***
## Residuals   133   5310    39.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test of region returns a p-value that is smaller than 2 x 10^(-16), very close to 0. We are confident to reject the null hypothesis of the ANOVA test. Thus, the differences in means of crude rate

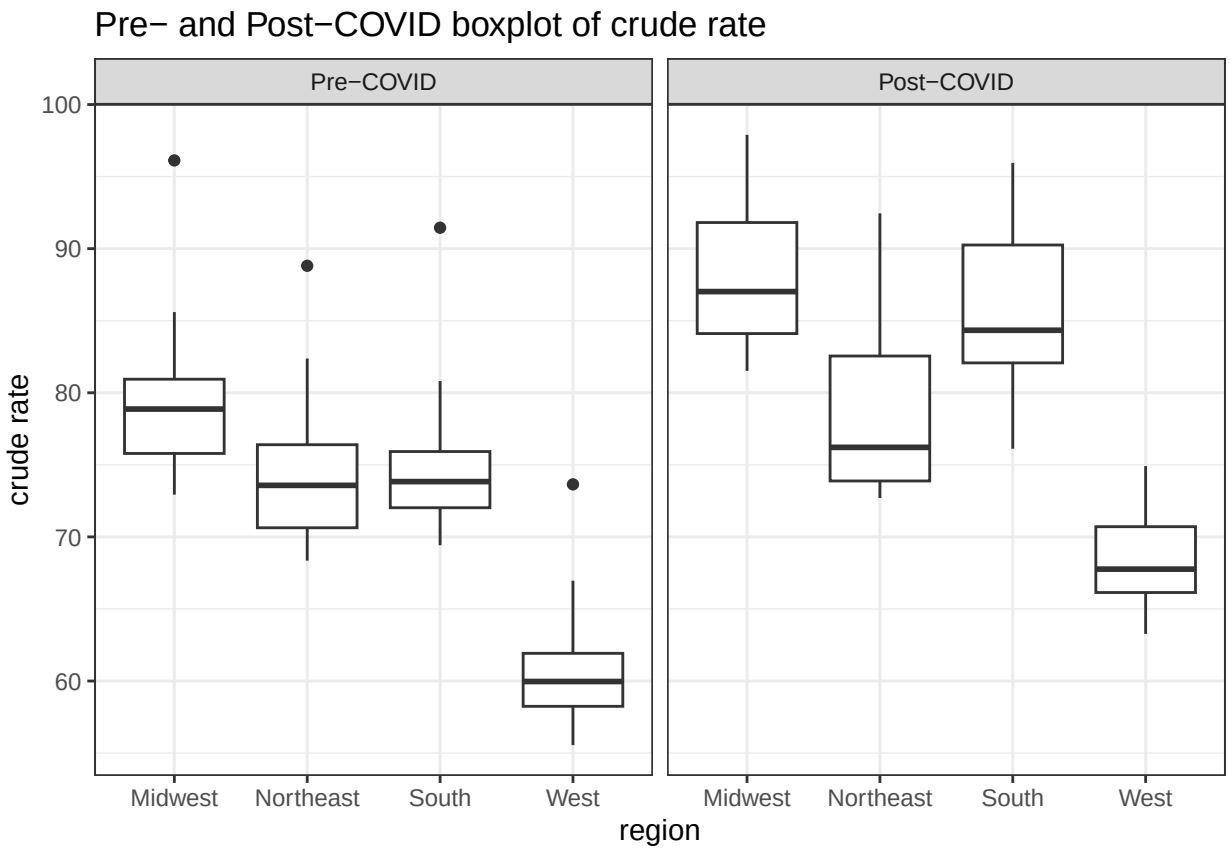
between each region are statistically significant.

Pre- vs. Post-COVID Analysis

In this section, we want to separate data into two groups: Pre- and Post-COVID. We want to study if there is statistical significant evidence that indicates the means of crude rate in each region are different.

General

```
region %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  ggplot(aes(x = region, y = crude_rate)) + geom_boxplot() +
  facet_wrap(~ factor(covid, levels = c("Pre-COVID", "Post-COVID")), nrow = 1) +
  labs(x = "region", y = "crude rate") +
  ggtitle("Pre- and Post-COVID boxplot of crude rate") +
  theme_bw()
```

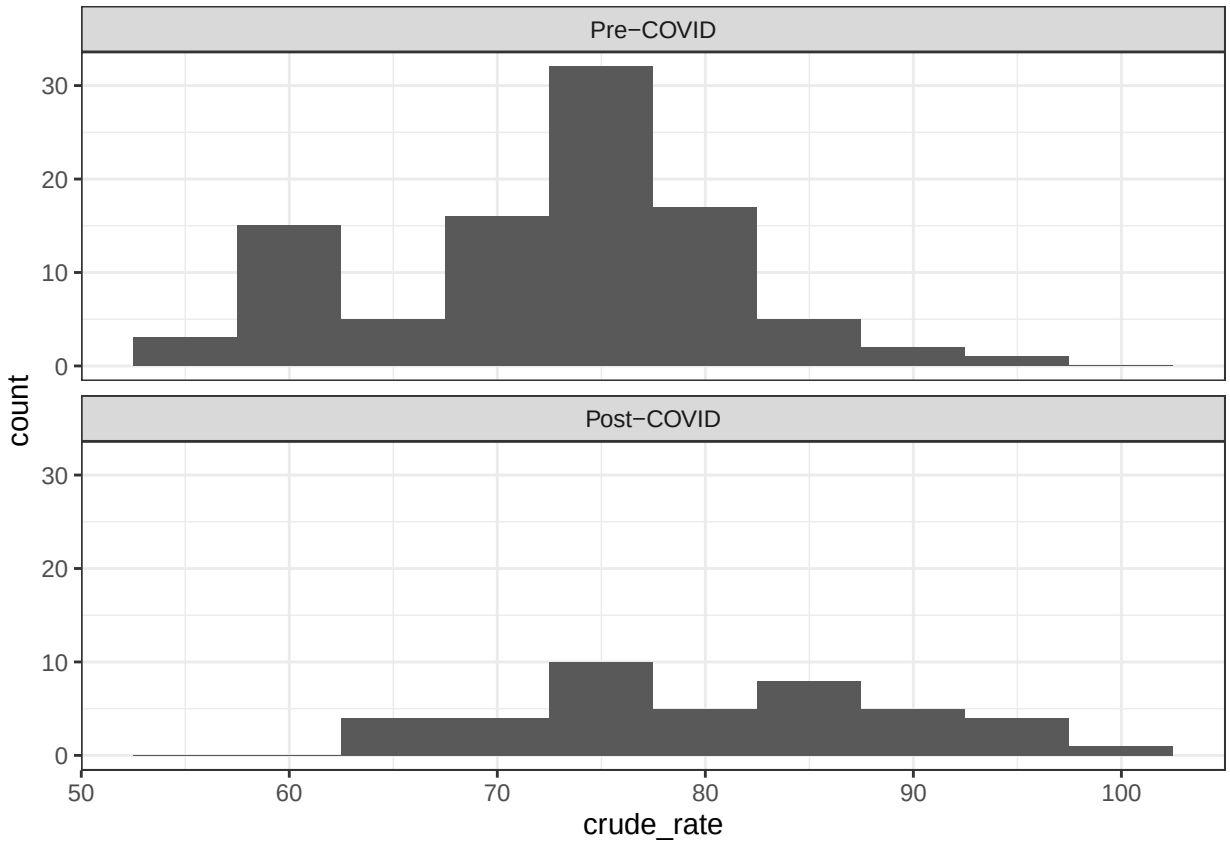



```
pre_post_crude_rate_mean <- region %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  group_by(covid) %>%
  summarize(mean_value = mean(crude_rate))
pre_post_crude_rate_mean
```

```
## # A tibble: 2 x 2
##   covid      mean_value
##   <chr>           <dbl>
## 1 Post-COVID       80.2
```

```
## 2 Pre-COVID        72.4
```

```
region %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  group_by(covid) %>% ggplot(aes(x=crude_rate)) + geom_histogram(binwidth = 5) +
  facet_wrap(~factor(covid, levels = c("Pre-COVID", "Post-COVID")), nrow = 2) +
  theme_bw()
```



The crude rate means Pre- and Post-COVID are different in general.

Each region

```
northeast <- region[region$region %in% "Northeast", ]
```

```
northeast %>%
  get_summary_stats(crude_rate, type = "mean_sd")
```

Northeast

```
## # A tibble: 1 x 4
##   variable       n  mean    sd
##   <fct>      <dbl> <dbl> <dbl>
## 1 crude_rate    33  75.6  5.80
```

```
northeast %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  group_by(covid) %>%
```

```
  summarise(var = var(crude_rate))
```

```
## # A tibble: 2 x 2
##   covid        var
##   <chr>      <dbl>
## 1 Post-COVID  46.8
## 2 Pre-COVID   25.4
```

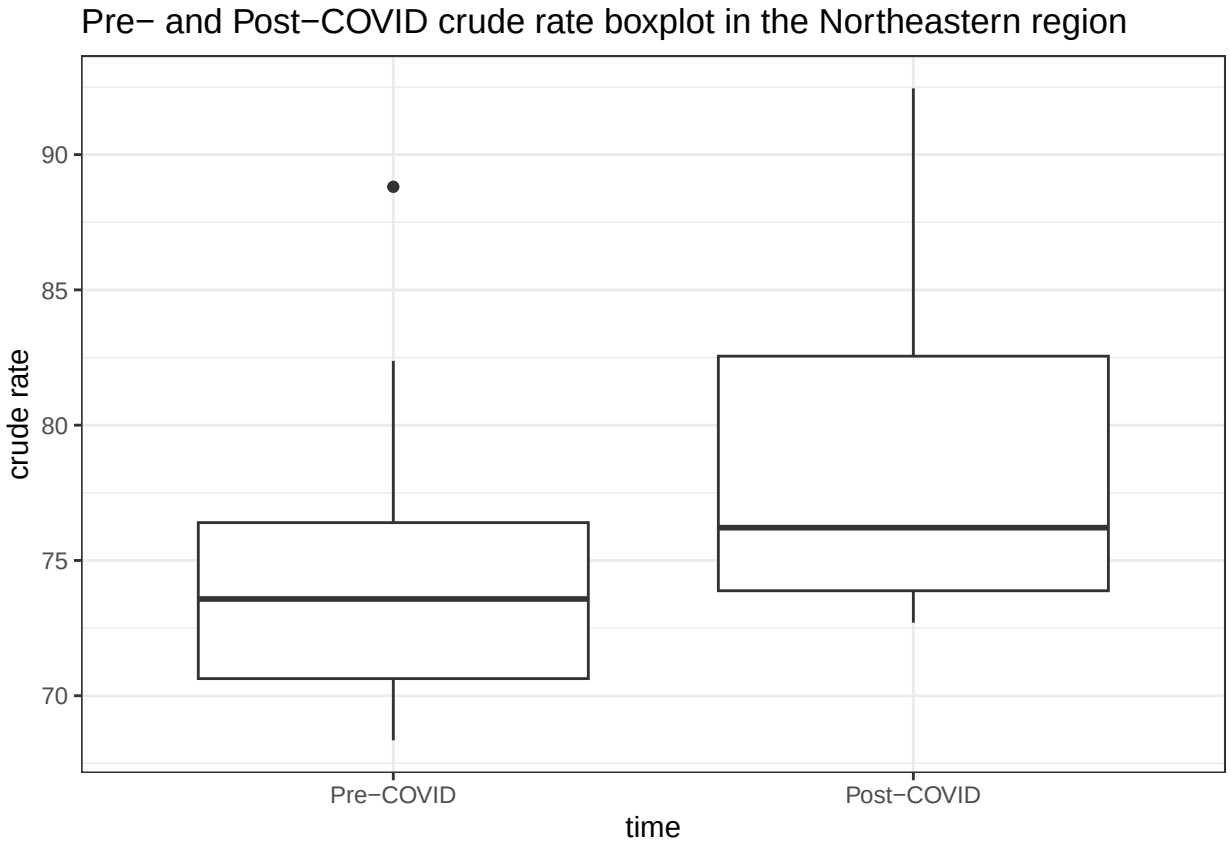
Because the variances of crude rate Pre- and Post-COVID are not close, we will set the parameter `var.equal` = FALSE.

```
northeast %>%
  t_test(crude_rate ~ covid, alternative = "less", var.equal = FALSE) %>%
  add_significance()
```

```
## # A tibble: 1 x 9
##   .y.        group1 group2    n1    n2 statistic    df      p p.signif
##   <chr>      <chr>  <chr>  <int> <int>     <dbl> <dbl>  <dbl> <chr>
## 1 crude_rate 0      1         24     9     -1.69  11.4 0.0591 ns
```

p-value of this test is 0.0591, which fails to reject the null hypothesis. The crude rate means in the Northeastern region are statistically same between Pre- and Post-COVID.

```
northeast %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  ggplot(aes(x = factor(covid, levels = c("Pre-COVID", "Post-COVID")),
             y = crude_rate)) + geom_boxplot() +
  labs(x = "time", y = "crude rate") +
  ggtitle("Pre- and Post-COVID crude rate boxplot in the Northeastern region") +
  theme_bw()
```



From the boxplot we observed that the means of crude rate Pre-COVID and Post-COVID are very close in the Northeastern region, which matched the conclusion from the t test.

```
midwest <- region[region$region %in% "Midwest", ]
```

```
midwest %>%
  get_summary_stats(crude_rate,type = "mean_sd")
```

Midwest

```
## # A tibble: 1 x 4
##   variable       n  mean    sd
##   <fct>      <dbl> <dbl> <dbl>
## 1 crude_rate    34  81.9  6.57
```

```
midwest %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  group_by(covid) %>%
  summarise(var = var(crude_rate))
```

```
## # A tibble: 2 x 2
##   covid        var
##   <chr>      <dbl>
## 1 Post-COVID  32.3
## 2 Pre-COVID   25.7
```

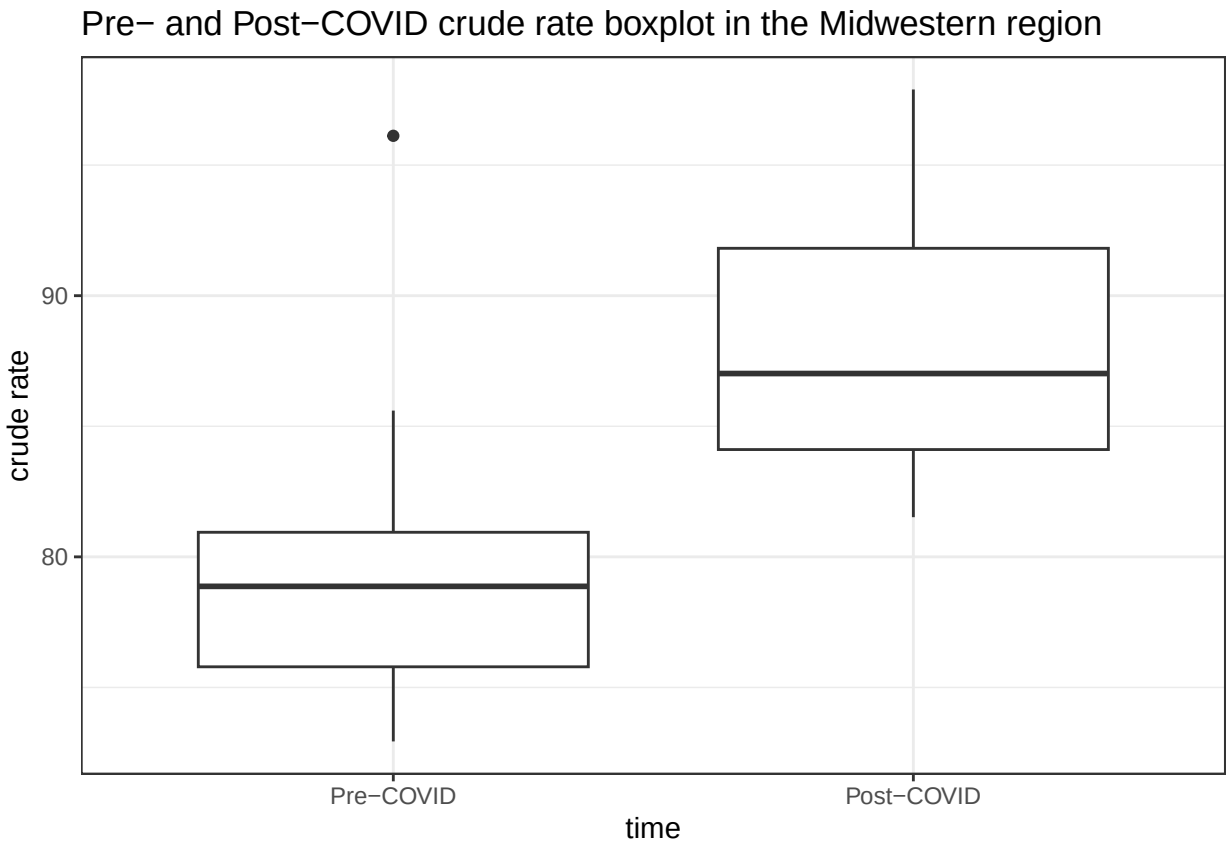
Because the variances of crude rate Pre- and Post-COVID are not close, we will set the parameter `var.equal` = FALSE.

```
midwest %>%
  t_test(crude_rate ~ covid, alternative = "less", var.equal = FALSE) %>%
  add_significance()
```

```
## # A tibble: 1 x 9
##   .y.        group1 group2    n1    n2 statistic    df        p p.signif
##   <chr>      <chr>  <chr>  <int> <int>     <dbl> <dbl>    <dbl> <chr>
## 1 crude_rate 0      1         24    10     -4.22  15.3 0.000359 ***
```

The p-value of the Midwestern Pre- and Post-COVID t test on crude rate means is 0.000359, which indicates that the null hypothesis should be rejected. The crude rate means are statistically different in the Midwestern region Pre- vs. Post-COVID.

```
midwest %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  ggplot(aes(x = factor(covid, levels = c("Pre-COVID", "Post-COVID")),
             y = crude_rate)) +
  geom_boxplot() + labs(x = "time", y = "crude rate") +
  ggtitle("Pre- and Post-COVID crude rate boxplot in the Midwestern region") +
  theme_bw()
```



From the boxplot, we observed the crude rate means are different Pre- vs. Post- COVID in the Midwestern region, which matched the conclusion from the t test.

```
south <- region[region$region %in% "South", ]
```

```
south %>%
  get_summary_stats(crude_rate,type = "mean_sd")
```

South

```
## # A tibble: 1 x 4
##   variable       n  mean    sd
##   <fct>      <dbl> <dbl> <dbl>
## 1 crude_rate    35  78.4  7.33
```

```
south %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  group_by(covid) %>%
  summarise(var = var(crude_rate))
```

```
## # A tibble: 2 x 2
##   covid        var
##   <chr>      <dbl>
## 1 Post-COVID  39.4
## 2 Pre-COVID   22.9
```

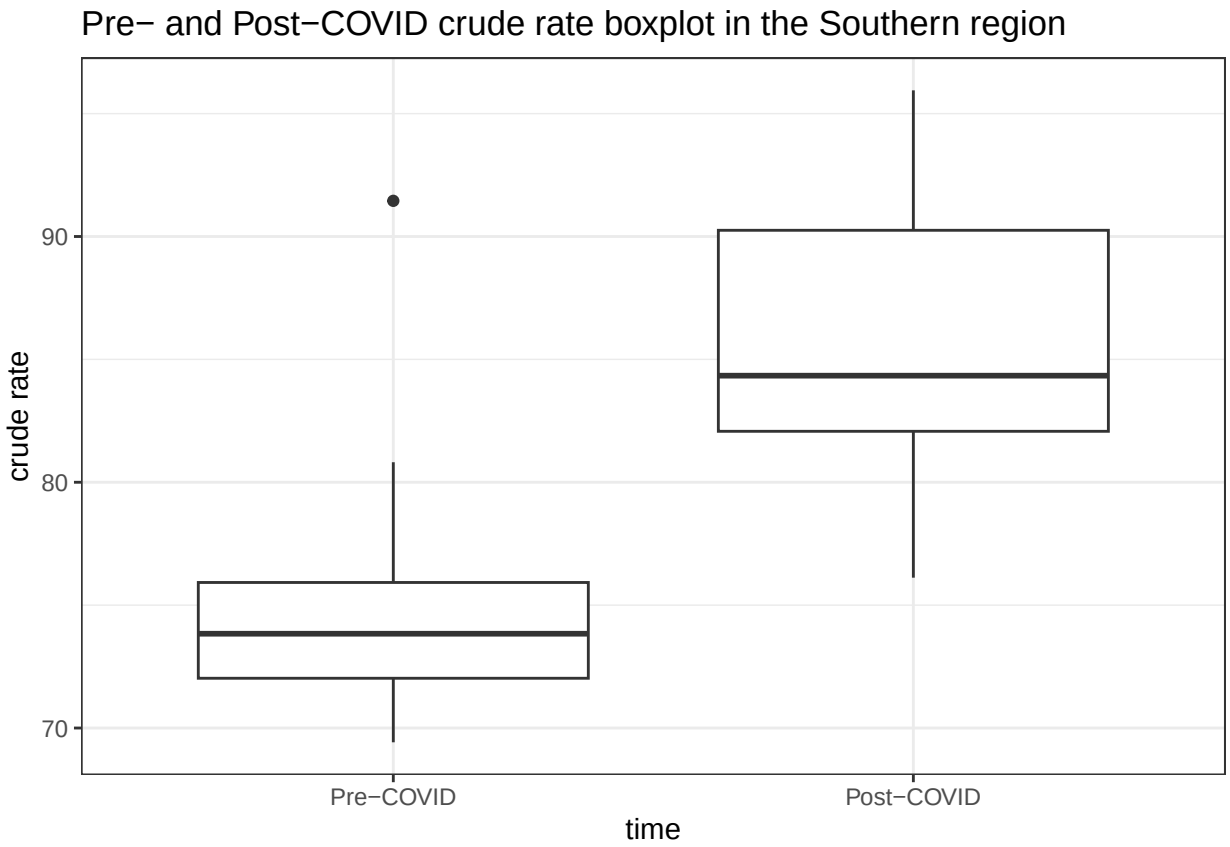
Because the variances of crude rate Pre- and Post-COVID are not close, we will set the parameter `var.equal` = FALSE.

```
south %>%
  t_test(crude_rate ~ covid, alternative = "less", var.equal = FALSE) %>%
  add_significance()
```

```
## # A tibble: 1 x 9
##   .y.        group1 group2    n1    n2 statistic    df         p p.signif
##   <chr>      <chr>  <chr>  <int> <int>     <dbl> <dbl>     <dbl> <chr>
## 1 crude_rate 0      1         24    11     -5.14  15.6 0.0000537 ****
```

The p-value of the Southern Pre- and Post-COVID t test on crude rate means is 5.37 x 10^(-5), which indicates that the null hypothesis should be rejected. The crude rate means are statistically different in the Southern region Pre- vs. Post-COVID.

```
south %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  ggplot(aes(x = factor(covid, levels = c("Pre-COVID", "Post-COVID")),
             y = crude_rate)) +
  geom_boxplot() + labs(x = "time", y = "crude rate") +
  ggtitle("Pre- and Post-COVID crude rate boxplot in the Southern region") +
  theme_bw()
```



From the boxplot, we observed the crude rate means are different Pre- vs. Post- COVID in the Southern region, which matched the conclusion from the t test.

```
west <- region[region$region %in% "West", ]
```

```
west %>%
  get_summary_stats(crude_rate,type = "mean_sd")
```

West

```
## # A tibble: 1 x 4
##   variable       n  mean    sd
##   <fct>      <dbl> <dbl> <dbl>
## 1 crude_rate    35  63.4  5.39
```

```
west %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  group_by(covid) %>%
  summarise(var = var(crude_rate))
```

```
## # A tibble: 2 x 2
##   covid        var
##   <chr>      <dbl>
## 1 Post-COVID  14.0
## 2 Pre-COVID   17.5
```

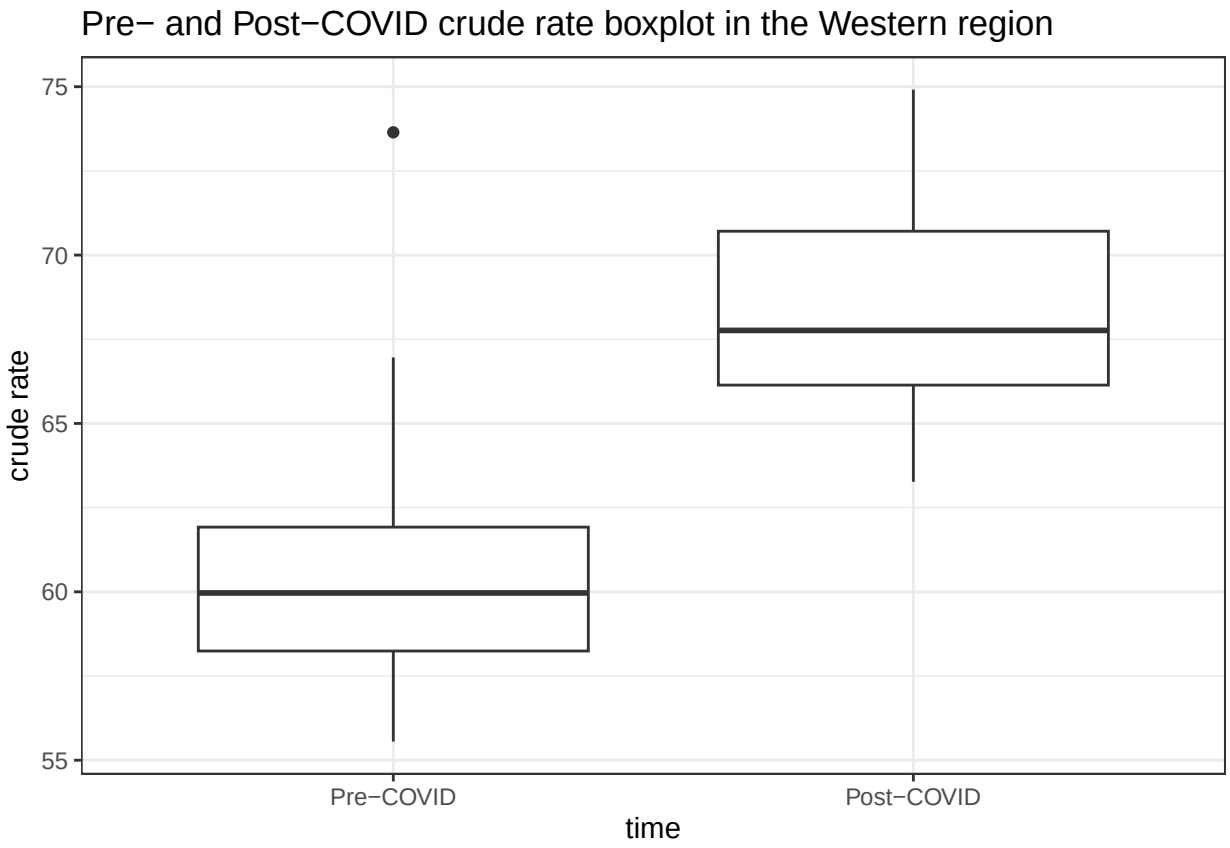
The variances are close. We will set `var.equal` = TRUE in the t test.

```
west %>%
  t_test(crude_rate ~ covid, alternative = "less", var.equal = TRUE) %>%
  add_significance()
```

```
## # A tibble: 1 x 9
##   .y.        group1 group2    n1    n2 statistic    df          p p.signif
##   <chr>      <chr>  <chr>  <int> <int>     <dbl> <dbl>      <dbl> <chr>
## 1 crude_rate 0      1         24    11     -5.21    33 0.00000497 ****
```

The p-value of the Western Pre- and Post-COVID t test on crude rate means is 4.97 x 10^(-6), which indicates that the null hypothesis should be rejected. The crude rate means are statistically different in the Western region Pre- vs. Post-COVID.

```
west %>%
  mutate(covid = recode(covid, "0" = "Pre-COVID", "1" = "Post-COVID")) %>%
  ggplot(aes(x = factor(covid, levels = c("Pre-COVID", "Post-COVID")),
             y = crude_rate)) +
  geom_boxplot() + labs(x = "time", y = "crude rate") +
  ggtitle("Pre- and Post-COVID crude rate boxplot in the Western region") +
  theme_bw()
```



From the boxplot, we observed the crude rate means are different Pre- vs. Post- COVID in the Southern region, which matched the conclusion from the t test.

Conclusion

1. The ANOVA result concludes that crude rate means are different among all regions.
2. The crude rate means Pre- and Post-COVID are different among all regions in general.
3. Midwestern, Soutern, Western regions display statistically significant difference in crude rate means Pre- and Post-COVID, while the Northeastern region does not.