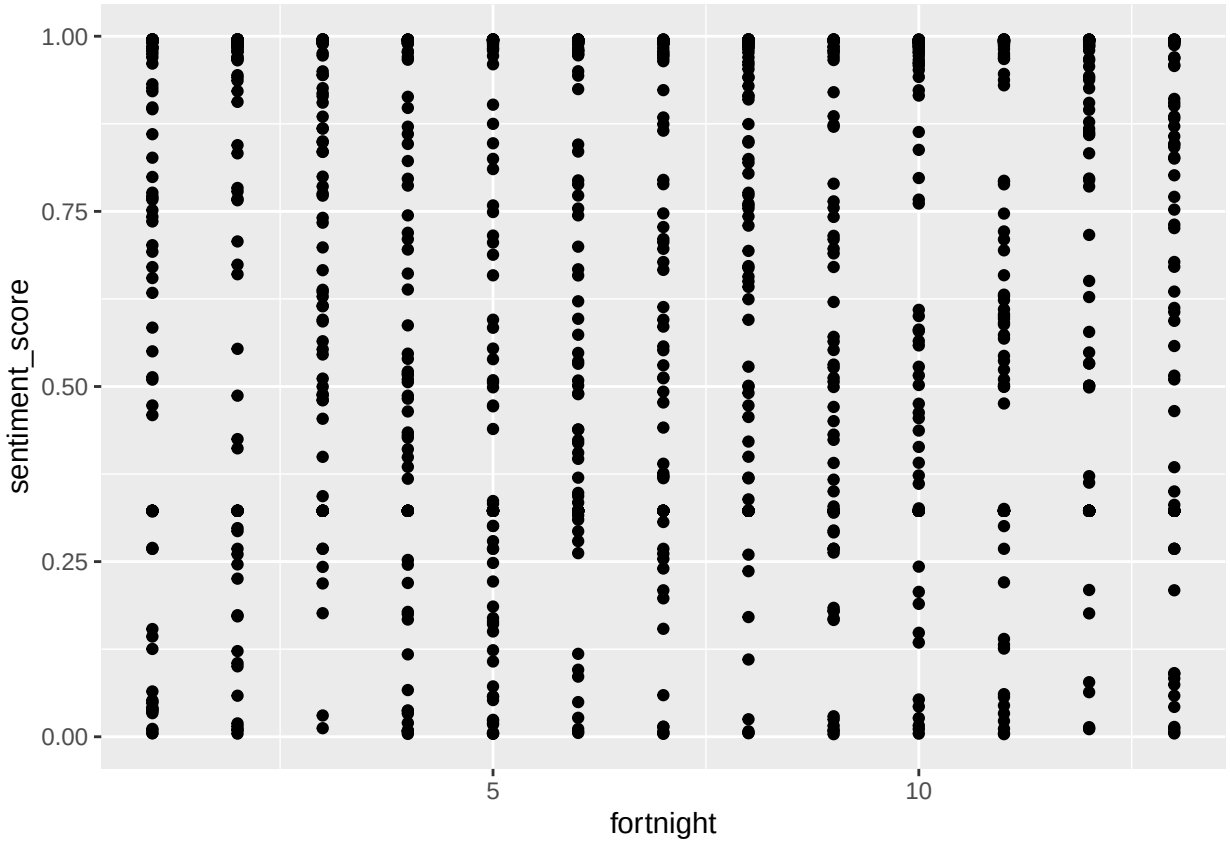
Datafest Data Analysis Canada

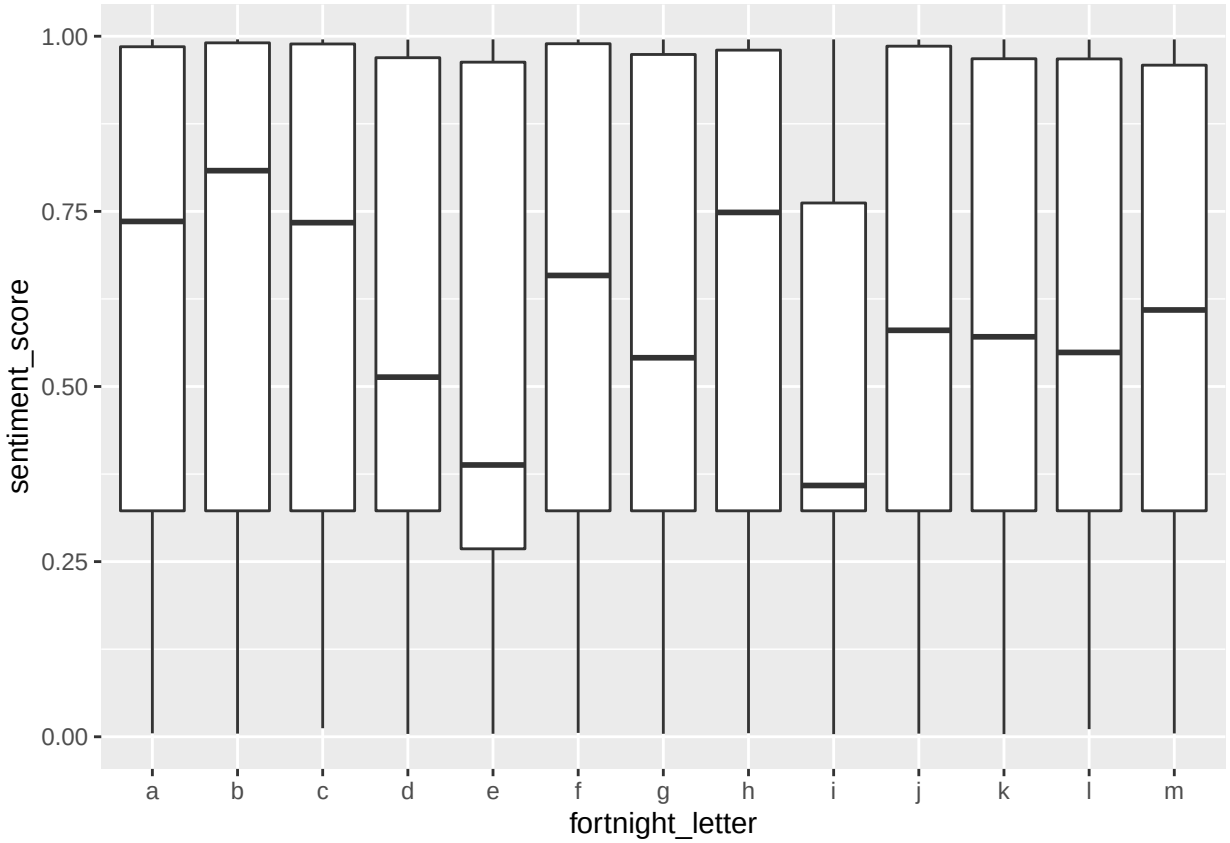
```
#US analysis import
Canada_analysis = read_excel("C:\\Users\\gtham\\OneDrive - Pomona College\\A - DATAFEST\\Analysis Datas
```

```
Canada_analysis_music <- Canada_analysis %>%
  filter(video_category == "Music")

Canada_analysis_travel <- Canada_analysis %>%
  filter(video_category == "Travel and Events")

Canada_analysis_people <- Canada_analysis %>%
  filter(video_category == "People and Blogs")

Canada_analysis_entertainment <- Canada_analysis %>%
  filter(video_category == "Entertainment")

Canada_analysis_news <- Canada_analysis %>%
  filter(video_category == "News and Politics")

Canada_analysis_how_to <- Canada_analysis %>%
  filter(video_category == "How-to and Style")

Canada_analysis_education <- Canada_analysis %>%
  filter(video_category == "Education")

Canada_analysis_science <- Canada_analysis %>%
  filter(video_category == "Science and Technology")
```

```
#full Canada data data summaries
ggplot(Canada_analysis) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



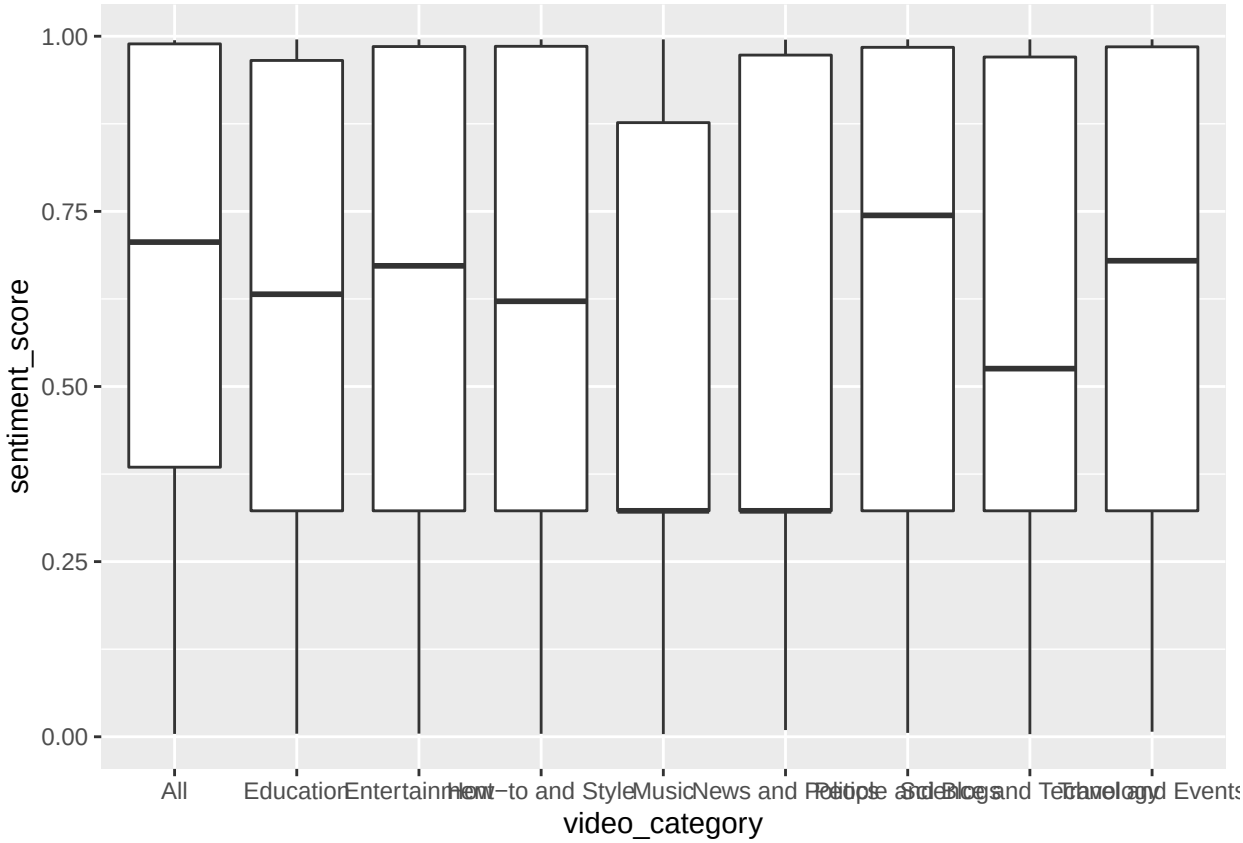

```
ggplot(Canada_analysis) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.611
##  2         2                   0.653
##  3         3                   0.672
##  4         4                   0.562
##  5         5                   0.518
##  6         6                   0.638
##  7         7                   0.586
##  8         8                   0.662
##  9         9                   0.511
## 10        10                   0.625
## 11        11                   0.572
## 12        12                   0.614
## 13        13                   0.593
```

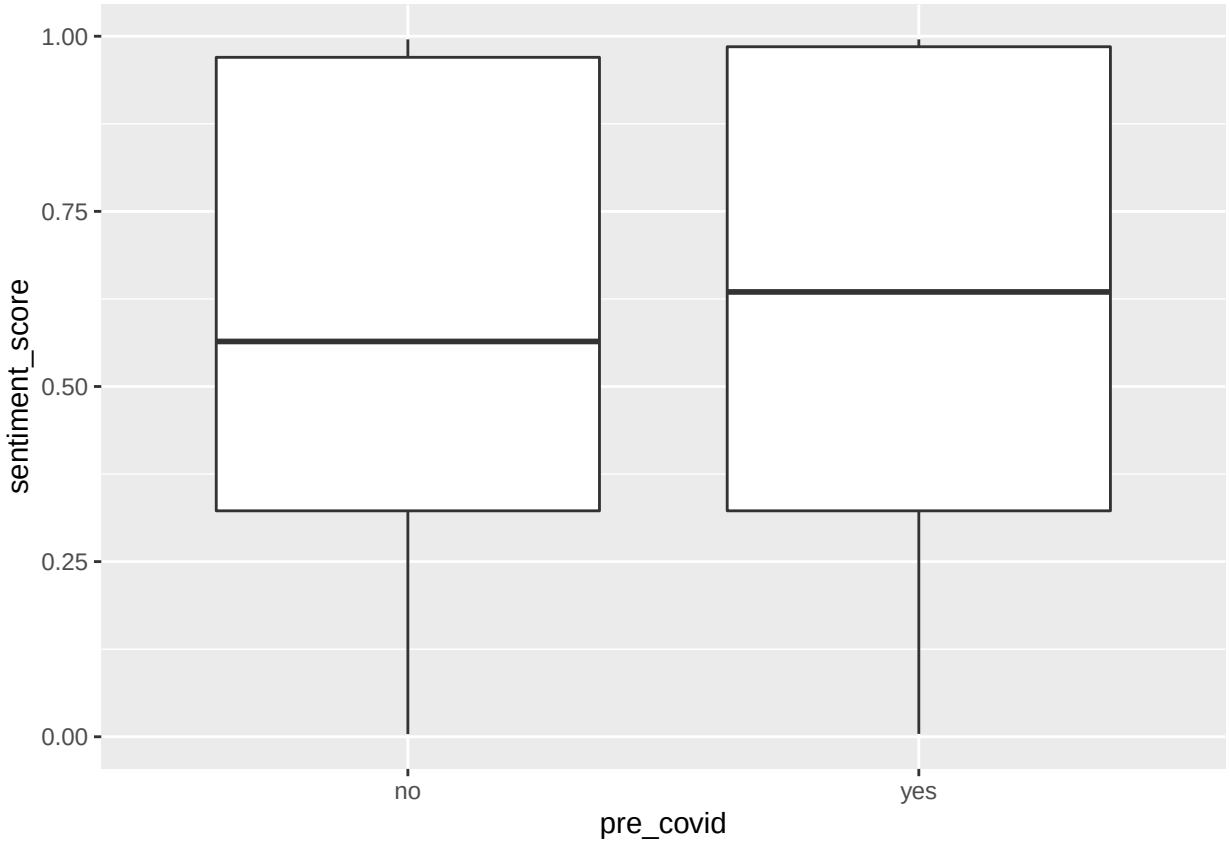
```
ggplot(Canada_analysis) +
  geom_boxplot(aes(x = video_category, y = sentiment_score))
```




```
Canada_analysis %>%
  group_by(video_category) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 9 x 2
##   video_category         `mean(sentiment_score)`
##   <chr>                                    <dbl>
## 1 All                                      0.641
## 2 Education                                0.603
## 3 Entertainment                            0.631
## 4 How-to and Style                         0.620
## 5 Music                                    0.474
## 6 News and Politics                        0.566
## 7 People and Blogs                         0.642
## 8 Science and Technology                   0.564
## 9 Travel and Events                        0.664
```

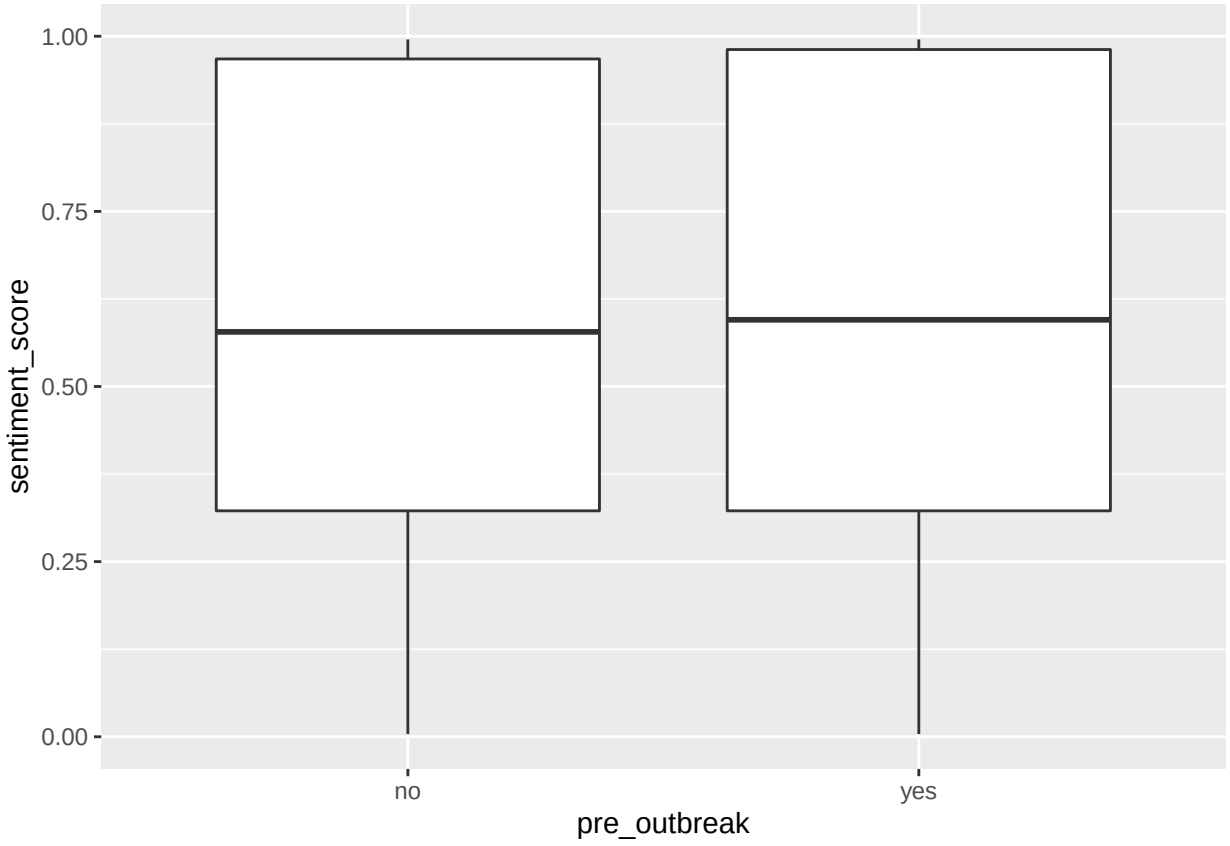
```
ggplot(Canada_analysis) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.595
## 2 yes                         0.609
```

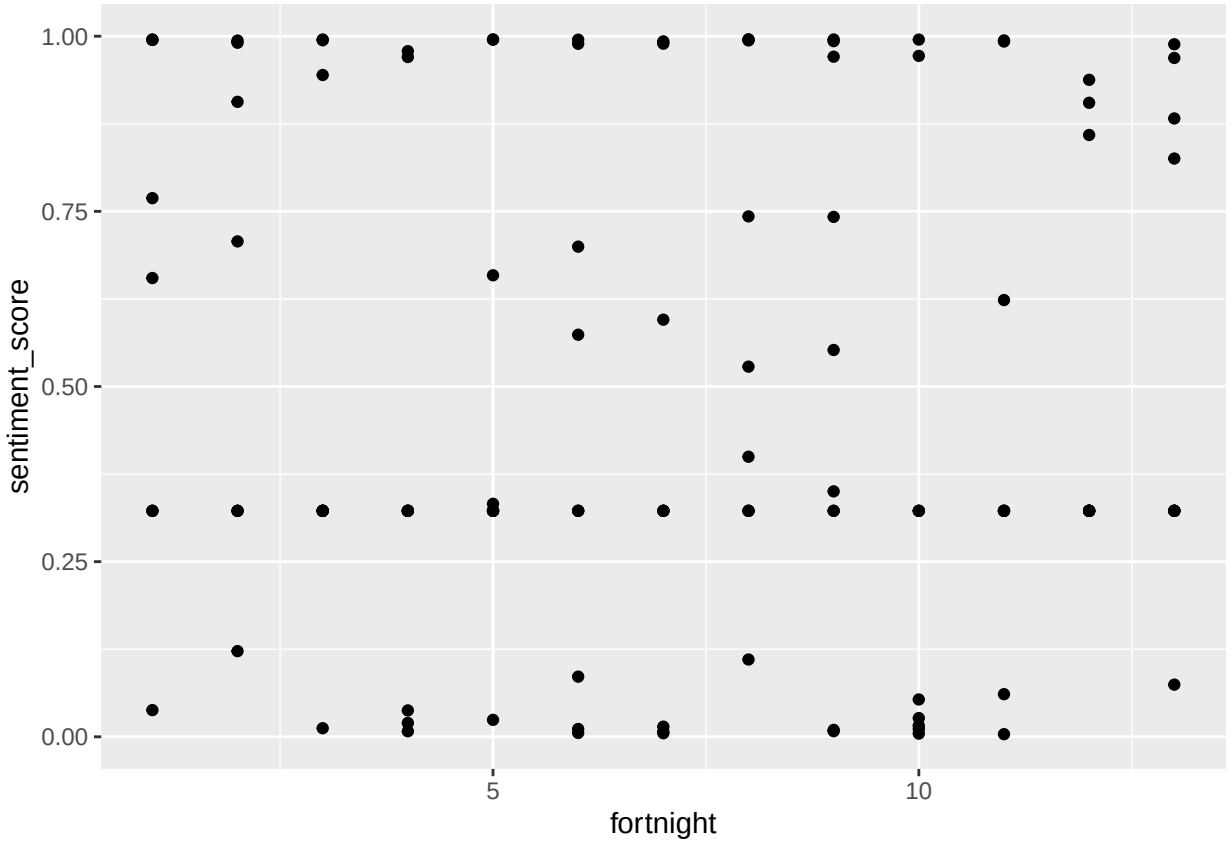
```
ggplot(Canada_analysis) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```



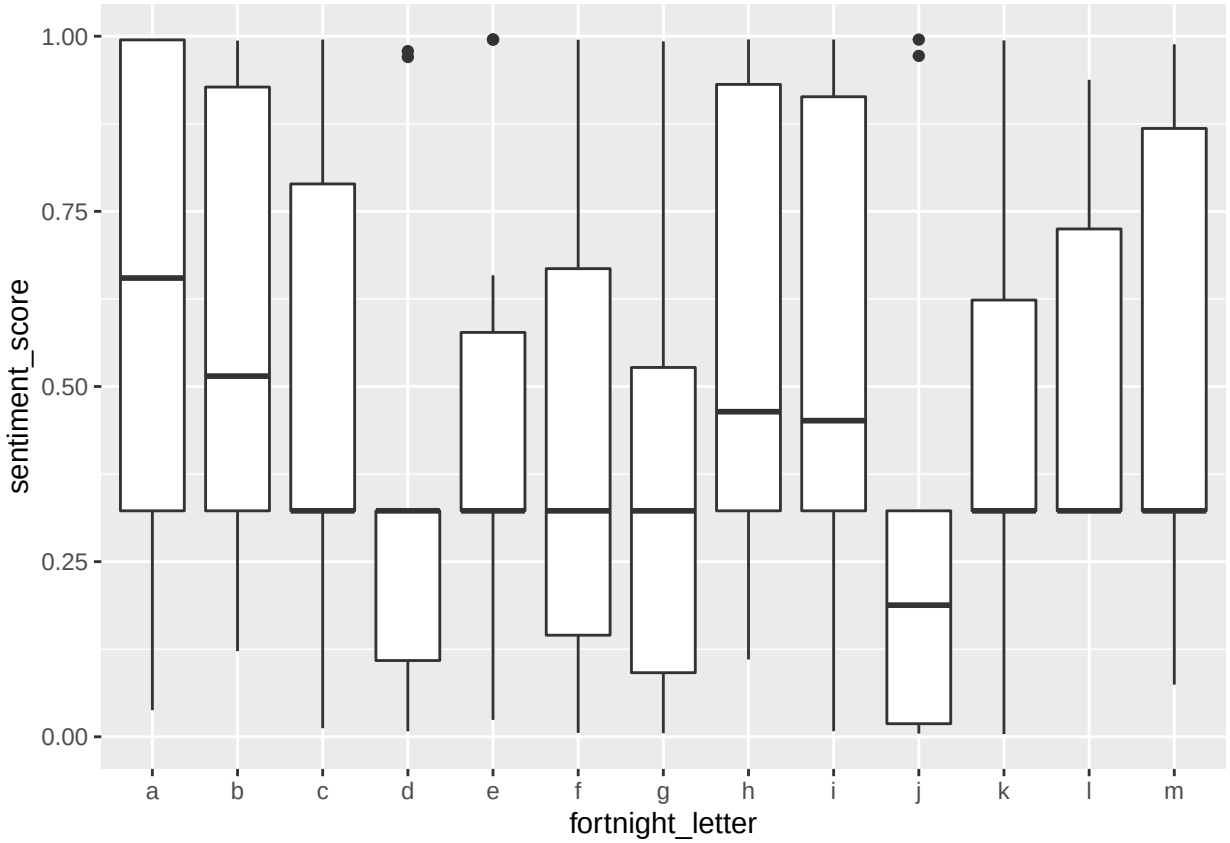

```
Canada_analysis %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.593
## 2 yes                            0.604
```

```
#data summary and analysis for music dataset
ggplot(Canada_analysis_music) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



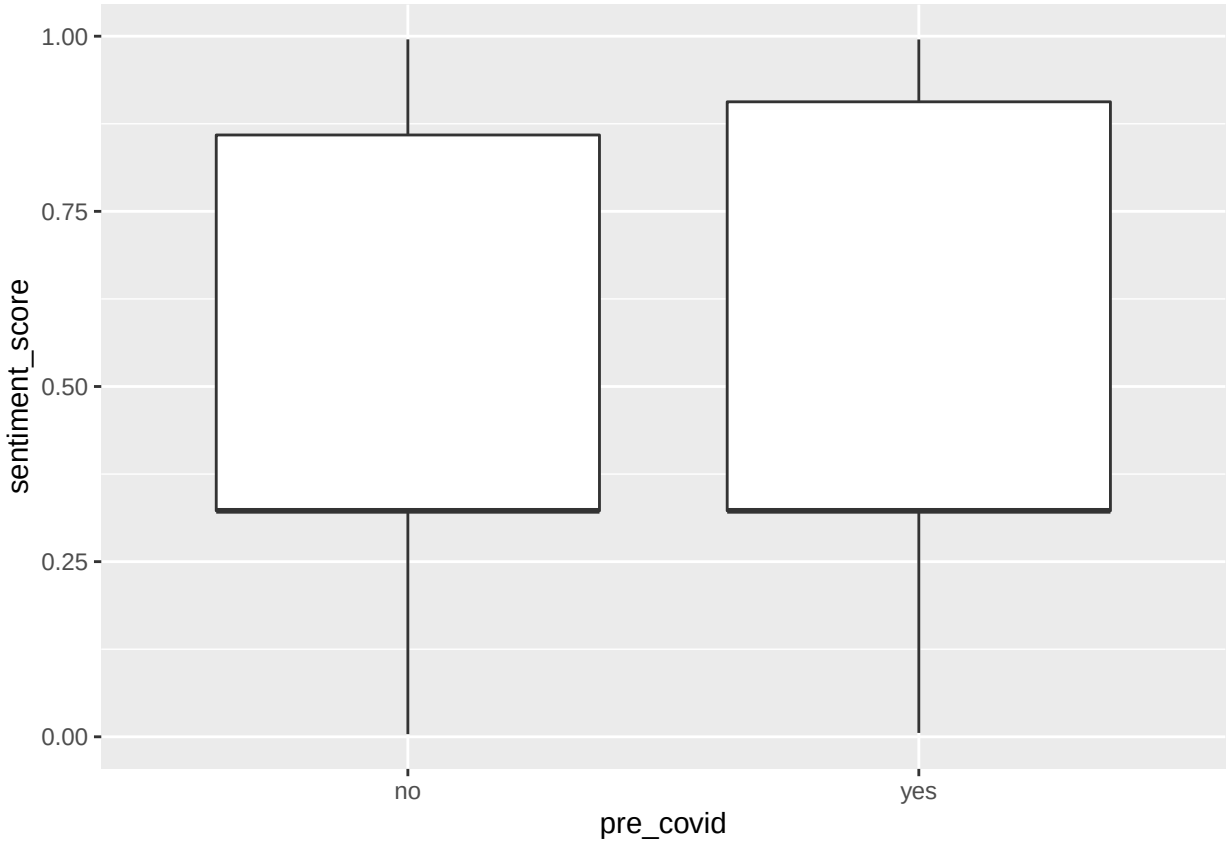

```
ggplot(Canada_analysis_music) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis_music %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.602
##  2         2                   0.586
##  3         3                   0.488
##  4         4                   0.363
##  5         5                   0.462
##  6         6                   0.433
##  7         7                   0.389
##  8         8                   0.573
##  9         9                   0.527
## 10        10                   0.305
## 11        11                   0.440
## 12        12                   0.496
## 13        13                   0.535
```

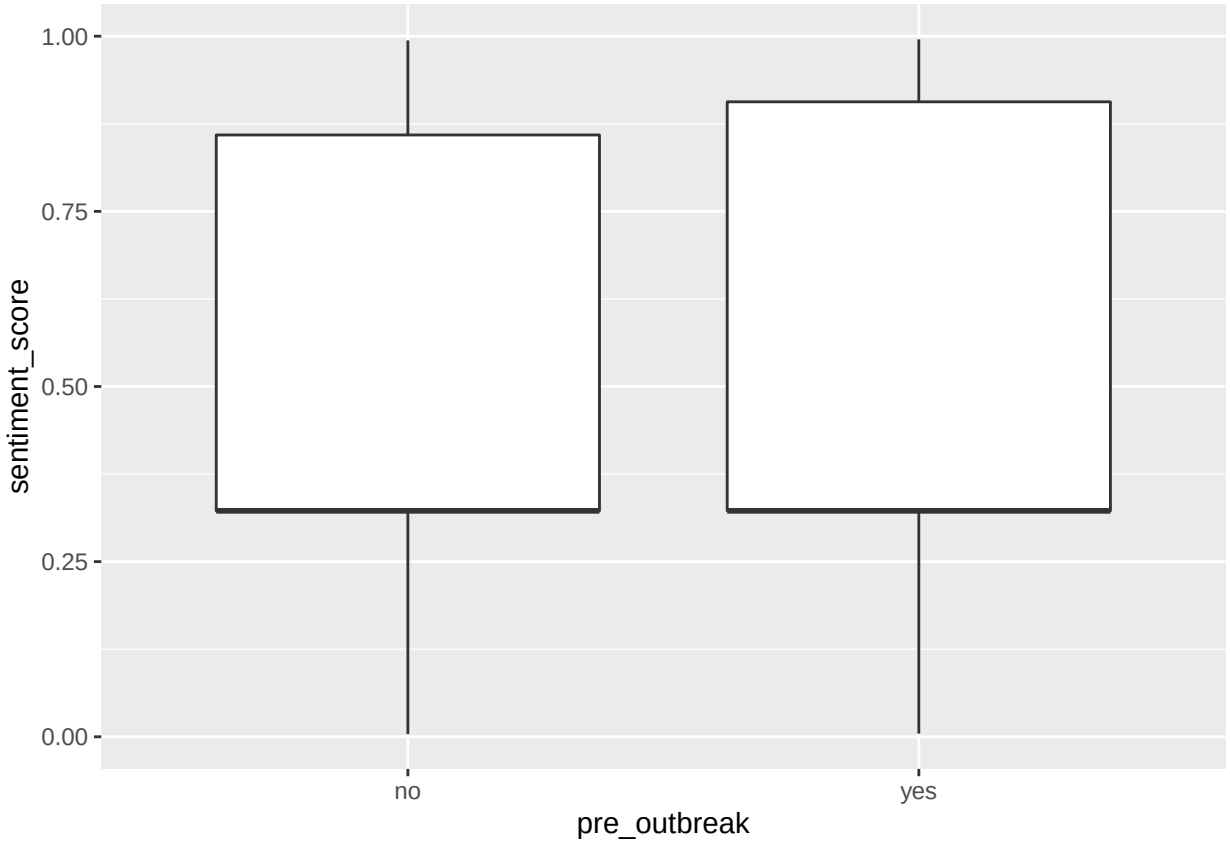
```
ggplot(Canada_analysis_music) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis_music %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.467
## 2 yes                         0.483
```

```
ggplot(Canada_analysis_music) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
Canada_analysis_music %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.492
## 2 yes                            0.469
```

```
#two proportion z-test for music dataset
#null hypothesis: the true proportion of positive sentiment music videos published precovid and postcov

count(Canada_analysis_music, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   69
## 2 TRUE                    57
```

```
m_num_precovid = 57
m_num_postcovid = 69
m_num = 126

Canada_analysis_music %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

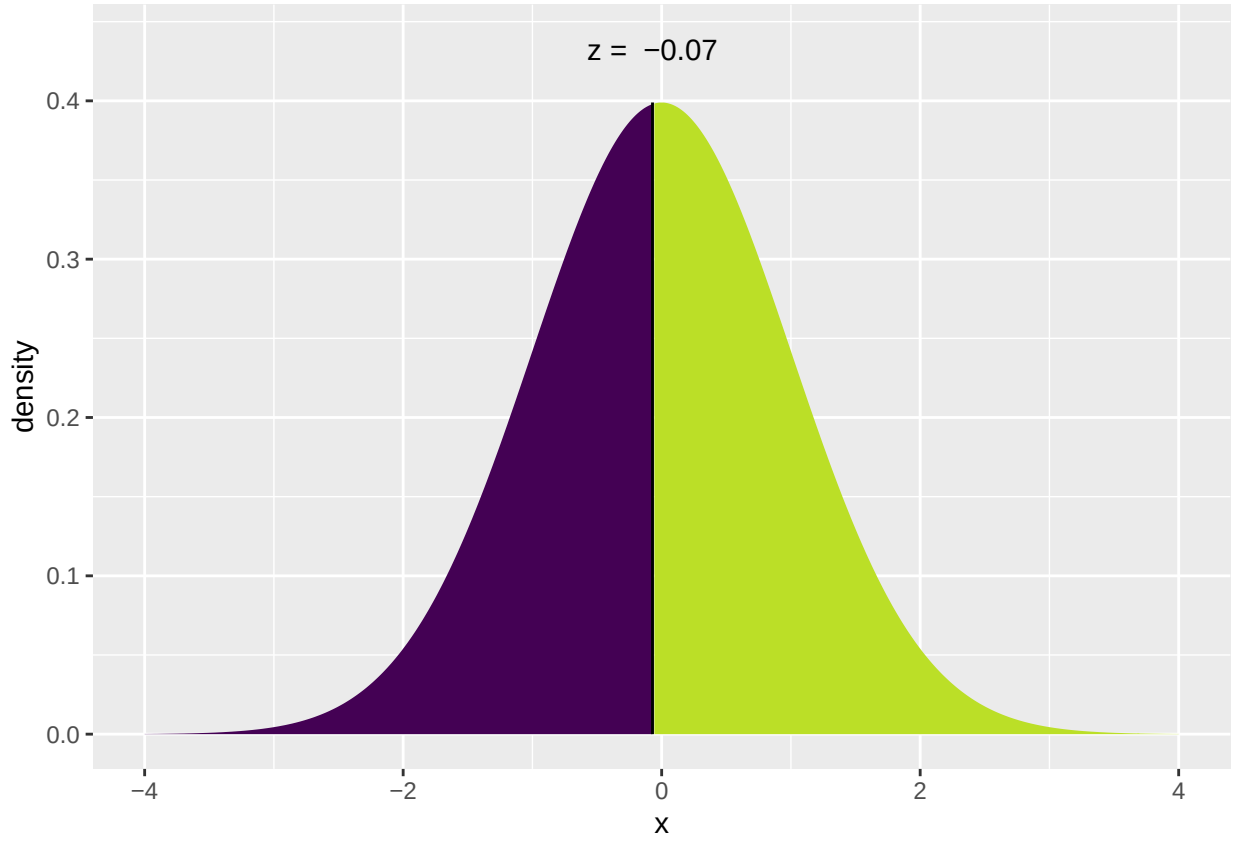
```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      36
## 2 TRUE                       21
```

```
p_hat_1_m_pos = 21/57

Canada_analysis_music %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      44
## 2 TRUE                       25
```

```
p_hat_2_m_pos = 25/69

p_hat_m_pos = (21+25)/(57+69)

sd <- sqrt((((p_hat_m_pos)*(1-p_hat_m_pos))/57)+(((p_hat_m_pos)*(1-p_hat_m_pos))/69))
z_score <- ((p_hat_2_m_pos-p_hat_1_m_pos)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.07081) = P(Z <= -0.07081) = 0.4718
##  P(X >  -0.07081) = P(Z >  -0.07081) = 0.5282
##
```




```
## [1] 0.9435469
```

```
#outbreak music
count(Canada_analysis_music, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      29
## 2 TRUE                       97
```

```
m_num_preoutbreak = 97
m_num_postoutbreak = 29
m_num = 126

Canada_analysis_music %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      61
## 2 TRUE                       36
```

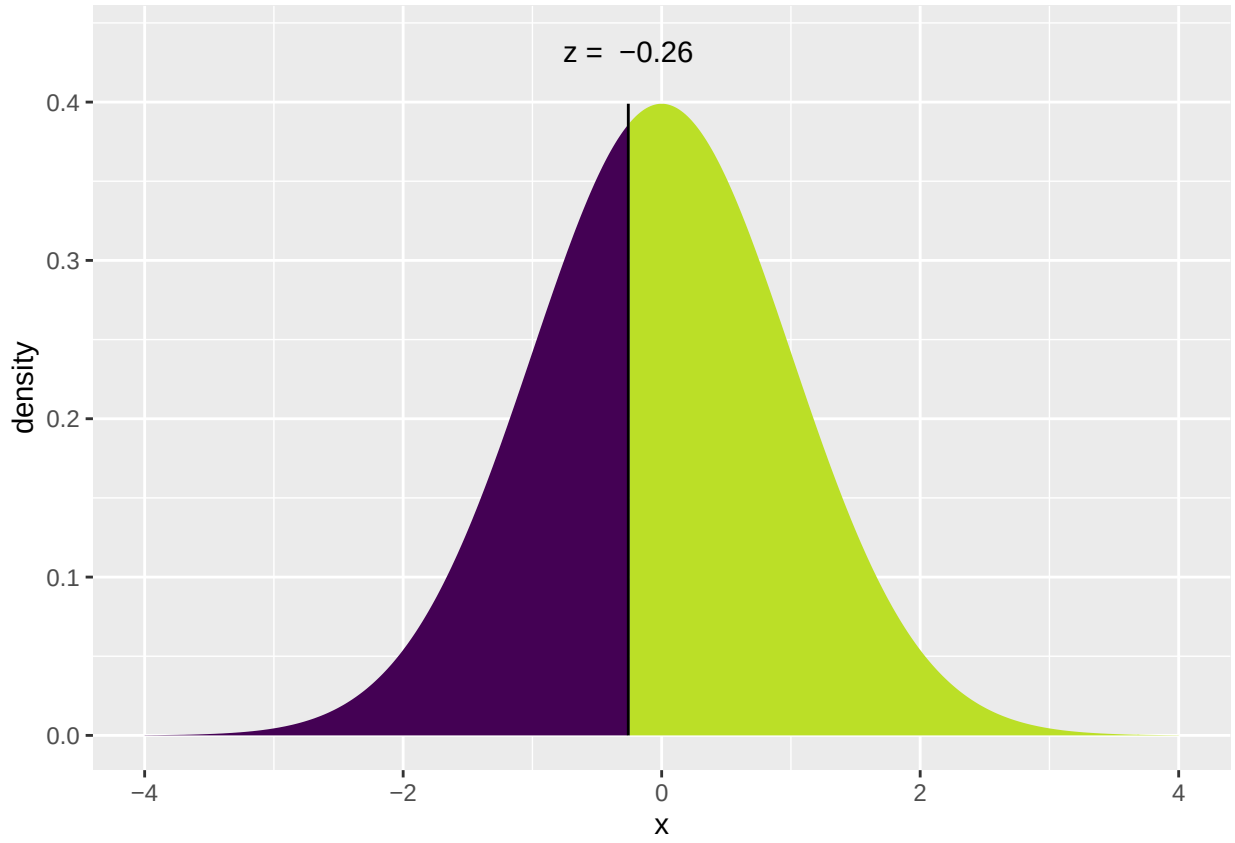
```
p_hat_1_m_pos = 36/97

Canada_analysis_music %>%
```

```
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      19
## 2 TRUE                       10
```

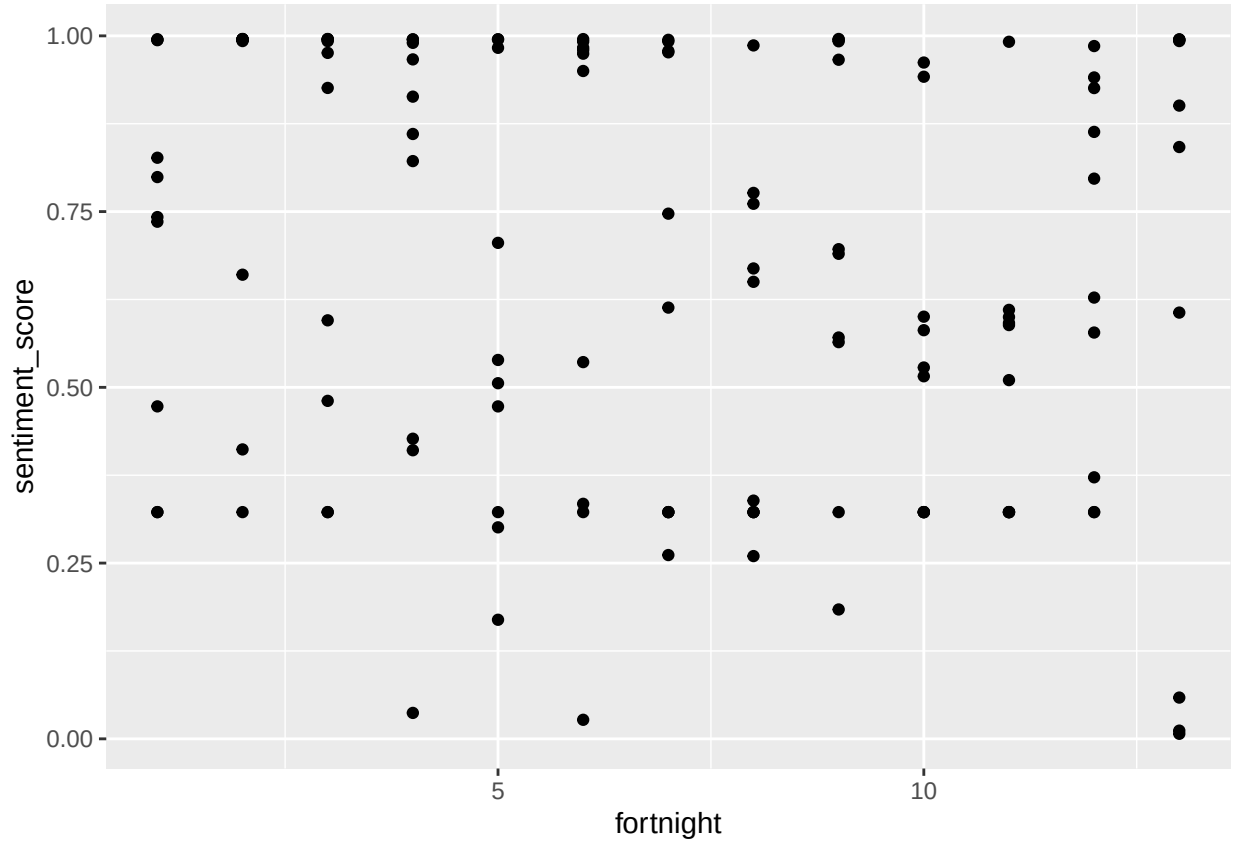
```
p_hat_2_m_pos = 10/29

p_hat_m_pos = (36+10)/(97+29)

sd <- sqrt((((p_hat_m_pos)*(1-p_hat_m_pos))/97)+(((p_hat_m_pos)*(1-p_hat_m_pos))/29))
z_score <- ((p_hat_2_m_pos-p_hat_1_m_pos)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.2582) = P(Z <= -0.2582) = 0.3981
##  P(X >  -0.2582) = P(Z >  -0.2582) = 0.6019
##
```

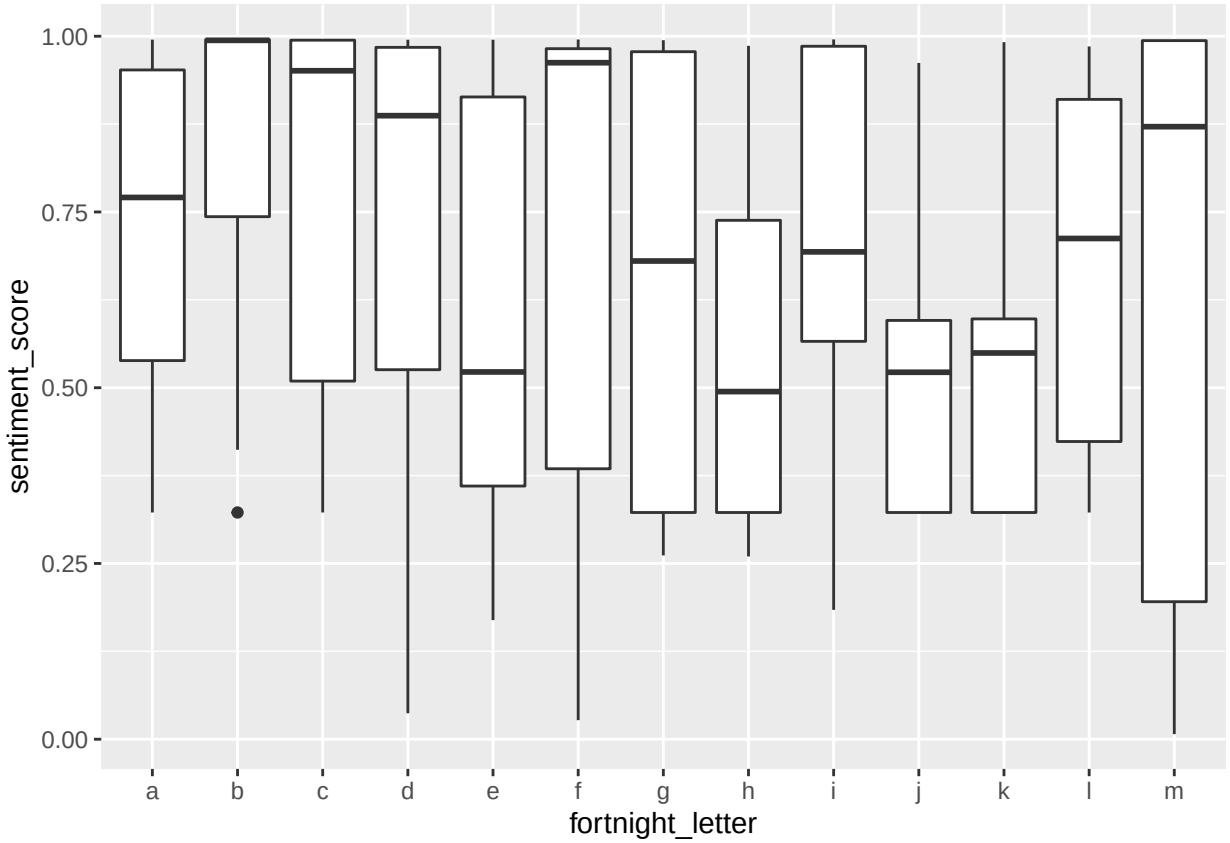


```
## [1] 0.7962746
```

```
#data summary travel
ggplot(Canada_analysis_travel) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



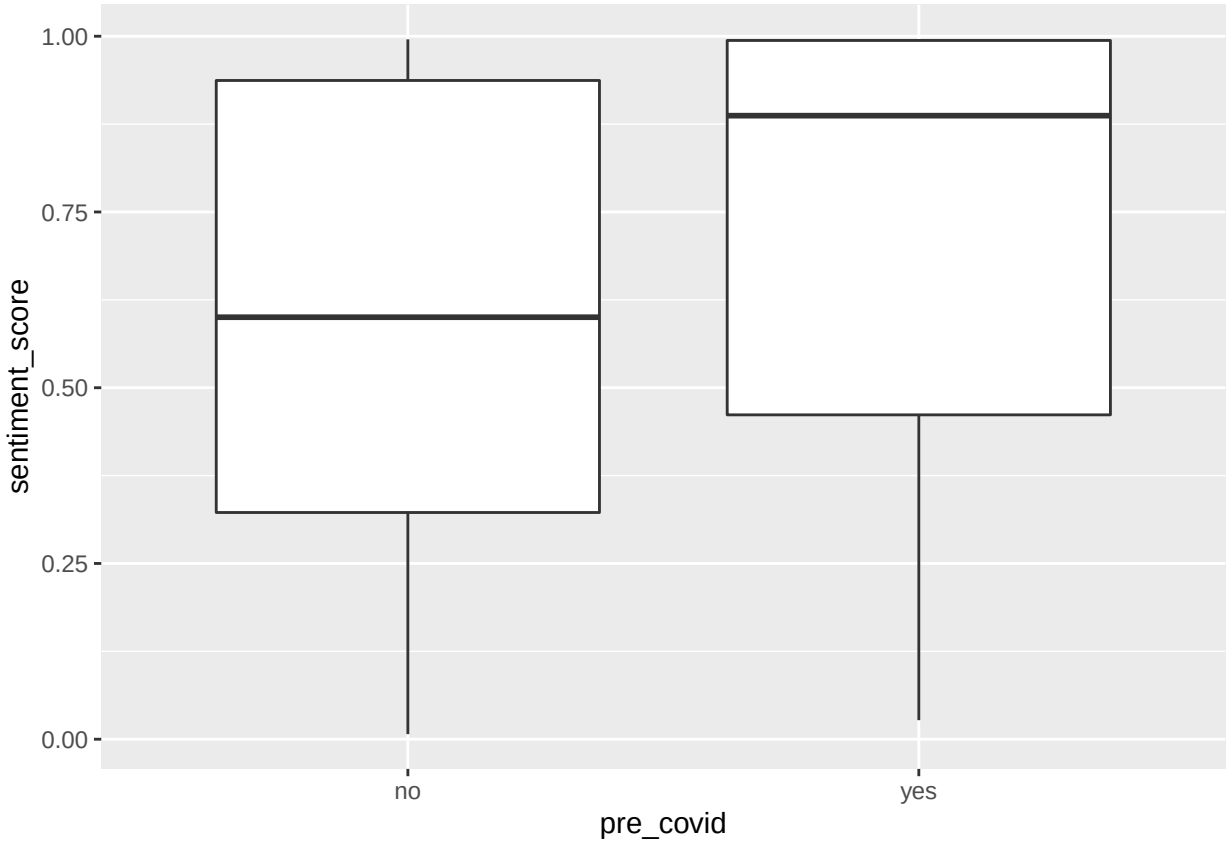

```
ggplot(Canada_analysis_travel) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis_travel %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.720
##  2         2                   0.836
##  3         3                   0.760
##  4         4                   0.742
##  5         5                   0.599
##  6         6                   0.709
##  7         7                   0.653
##  8         8                   0.541
##  9         9                   0.698
## 10        10                   0.542
## 11        11                   0.518
## 12        12                   0.674
## 13        13                   0.640
```

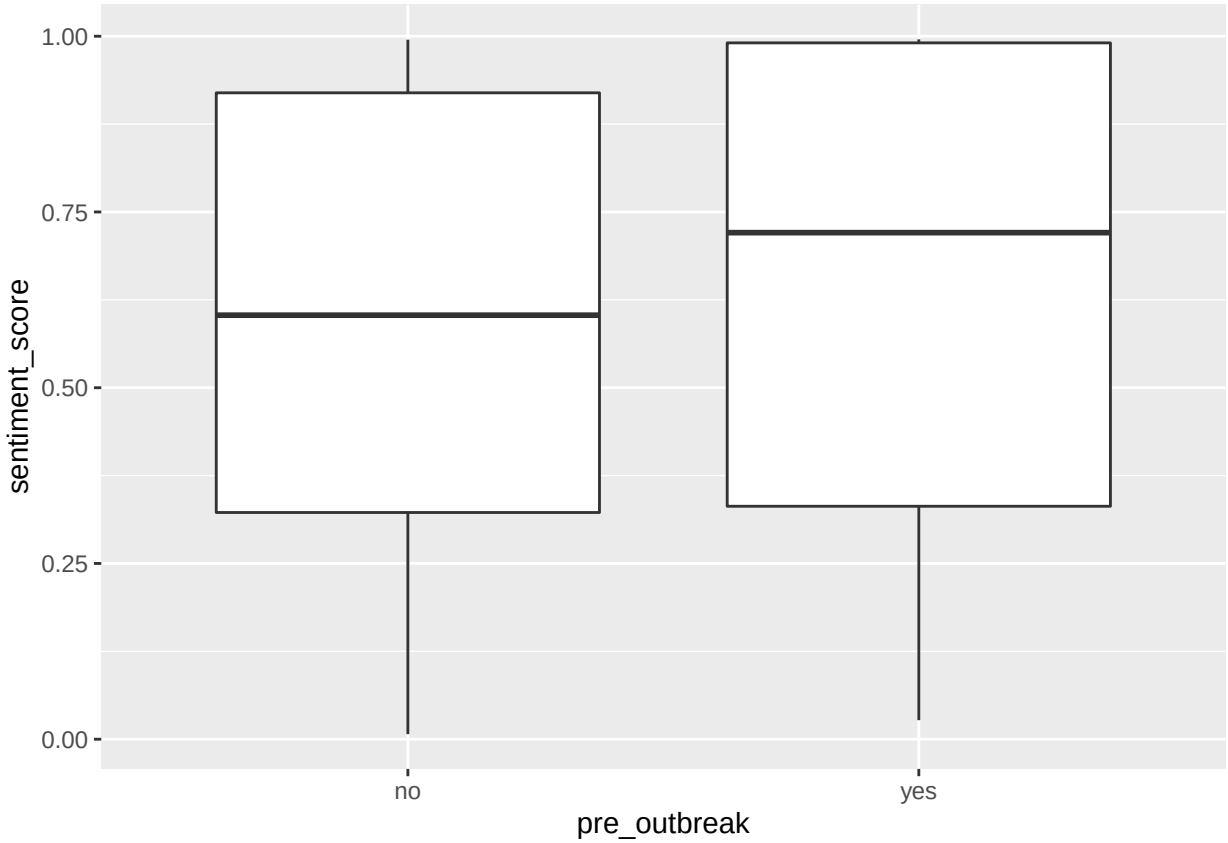
```
ggplot(Canada_analysis_travel) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis_travel %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.609
## 2 yes                         0.728
```

```
ggplot(Canada_analysis_travel) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
Canada_analysis_travel %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.611
## 2 yes                            0.680
```

```
#precovid travel
#null hypothesis: the true proportion of positive sentiment travel videos published precovid and postco

count(Canada_analysis_travel, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    60
```

```
t_num_precovid = 60
t_num_postcovid = 70
t_num = 130

Canada_analysis_travel %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

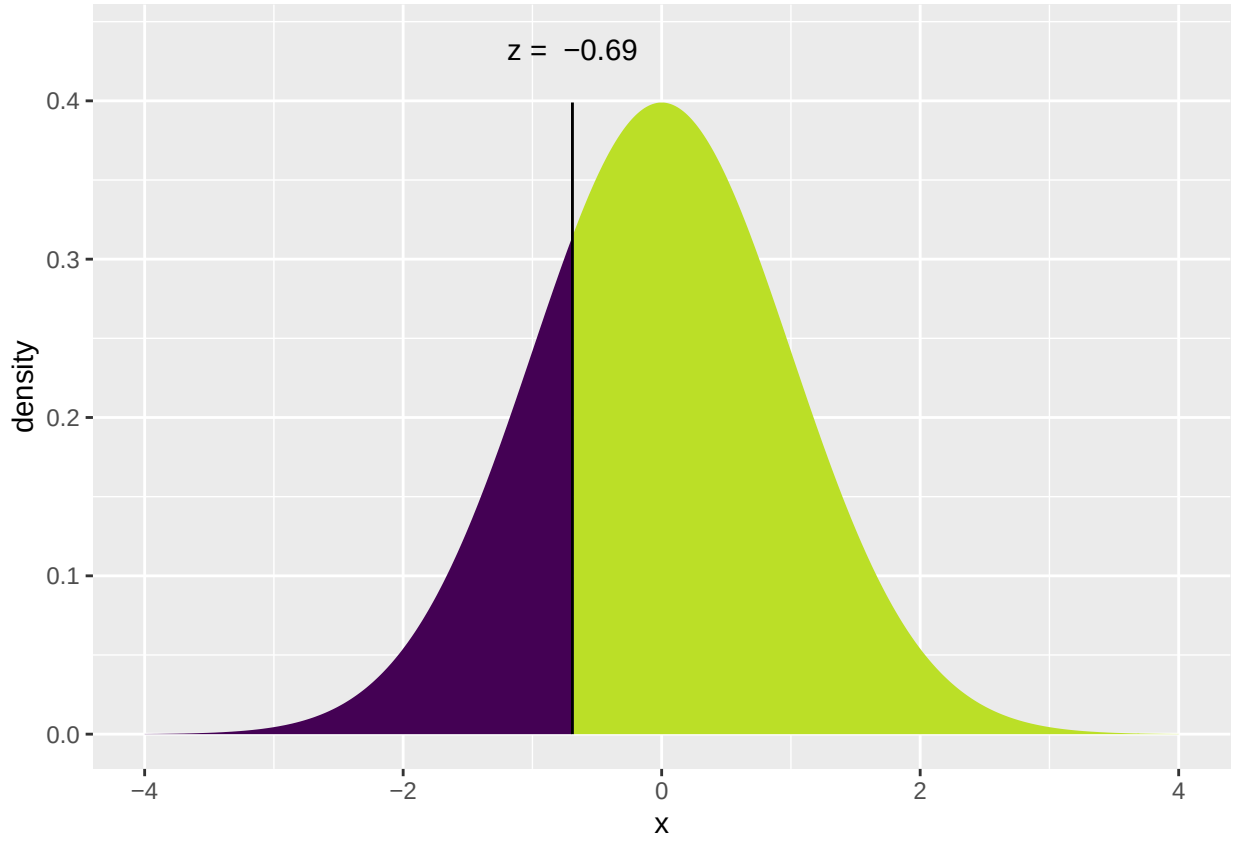
```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      18
## 2 TRUE                       42
```

```
p_hat_1_t_pos = 42/60

Canada_analysis_travel %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      25
## 2 TRUE                       45
```

```
p_hat_2_t_pos = 45/70

p_hat_t_pos = (42+45)/(60+70)

sd <- sqrt((((p_hat_t_pos)*(1-p_hat_t_pos))/60)+(((p_hat_t_pos)*(1-p_hat_t_pos))/70))
z_score <- ((p_hat_2_t_pos-p_hat_1_t_pos)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.6903) = P(Z <= -0.6903) = 0.245
##  P(X >  -0.6903) = P(Z >  -0.6903) = 0.755
##
```




```
## [1] 0.489979
```

```
#outbreak travel
count(Canada_analysis_travel, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                      100
```

```
t_num_preoutbreak = 100
t_num_postoutbreak = 30
t_num = 130

Canada_analysis_travel %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      33
## 2 TRUE                       67
```

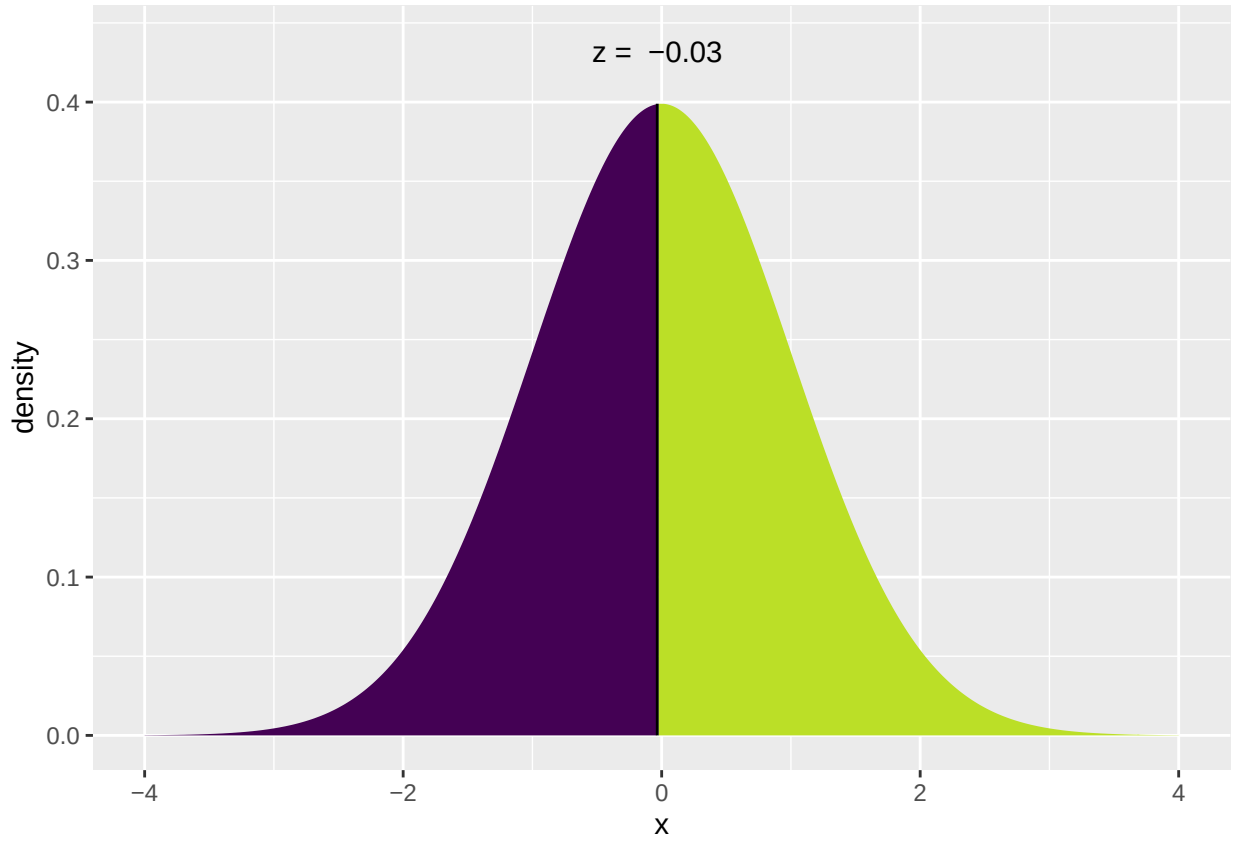
```
p_hat_1_t_pos = 67/100

Canada_analysis_travel %>%
```

```
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      10
## 2 TRUE                       20
```

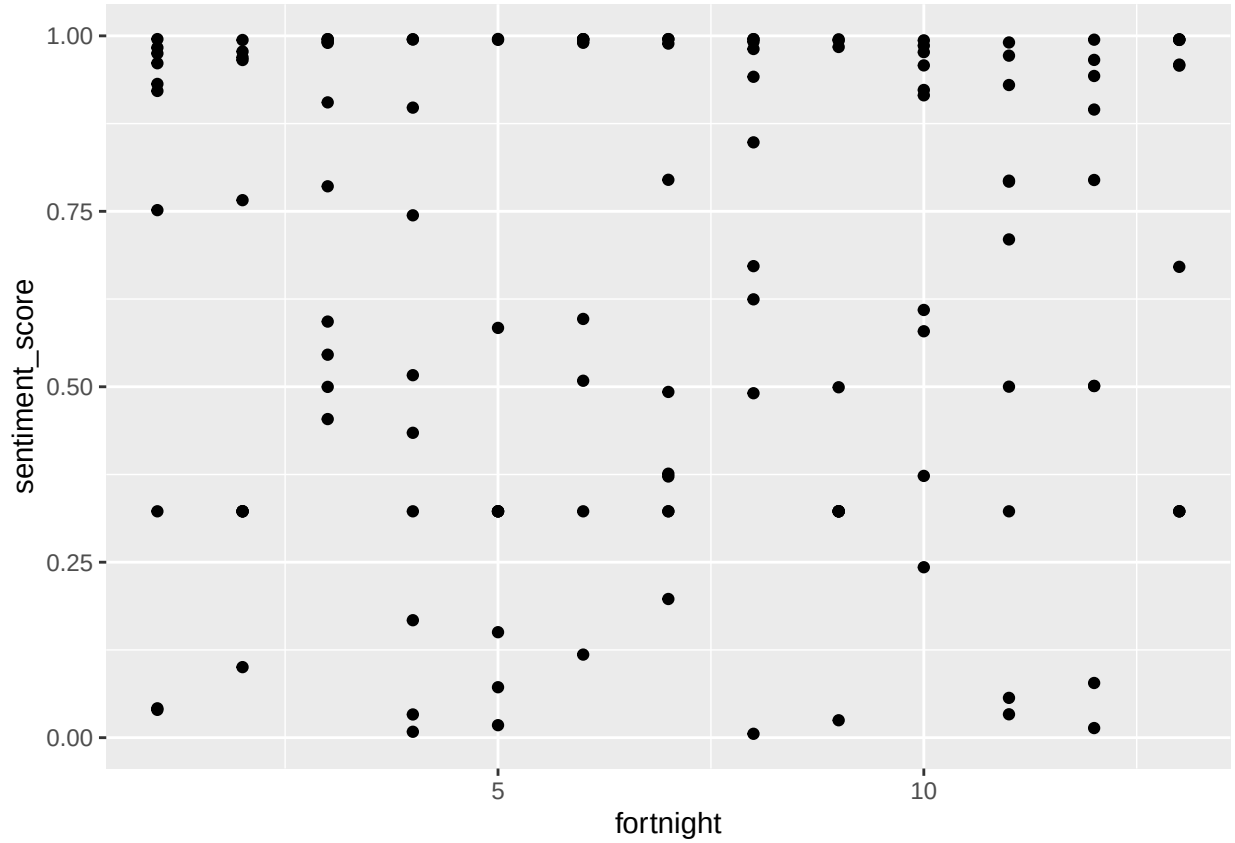
```
p_hat_2_t_pos = 20/30

p_hat_t_pos = (67+20)/(100+30)

sd <- sqrt((((p_hat_t_pos)*(1-p_hat_t_pos))/100)+(((p_hat_t_pos)*(1-p_hat_t_pos))/30))
z_score <- ((p_hat_2_t_pos-p_hat_1_t_pos)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.03403) = P(Z <= -0.03403) = 0.4864
##  P(X >  -0.03403) = P(Z >  -0.03403) = 0.5136
##
```

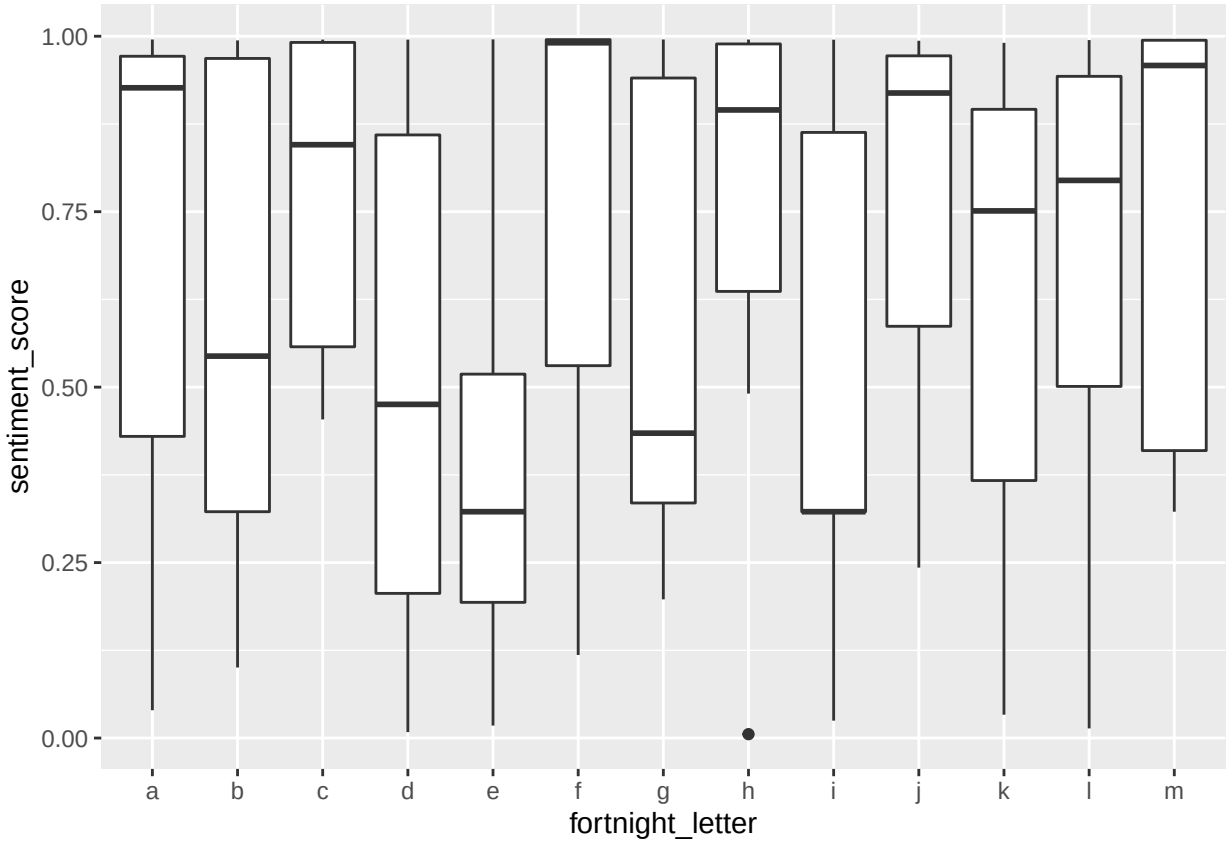


```
## [1] 0.9728498
```

```
#data summary people and blogs
ggplot(Canada_analysis_people) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



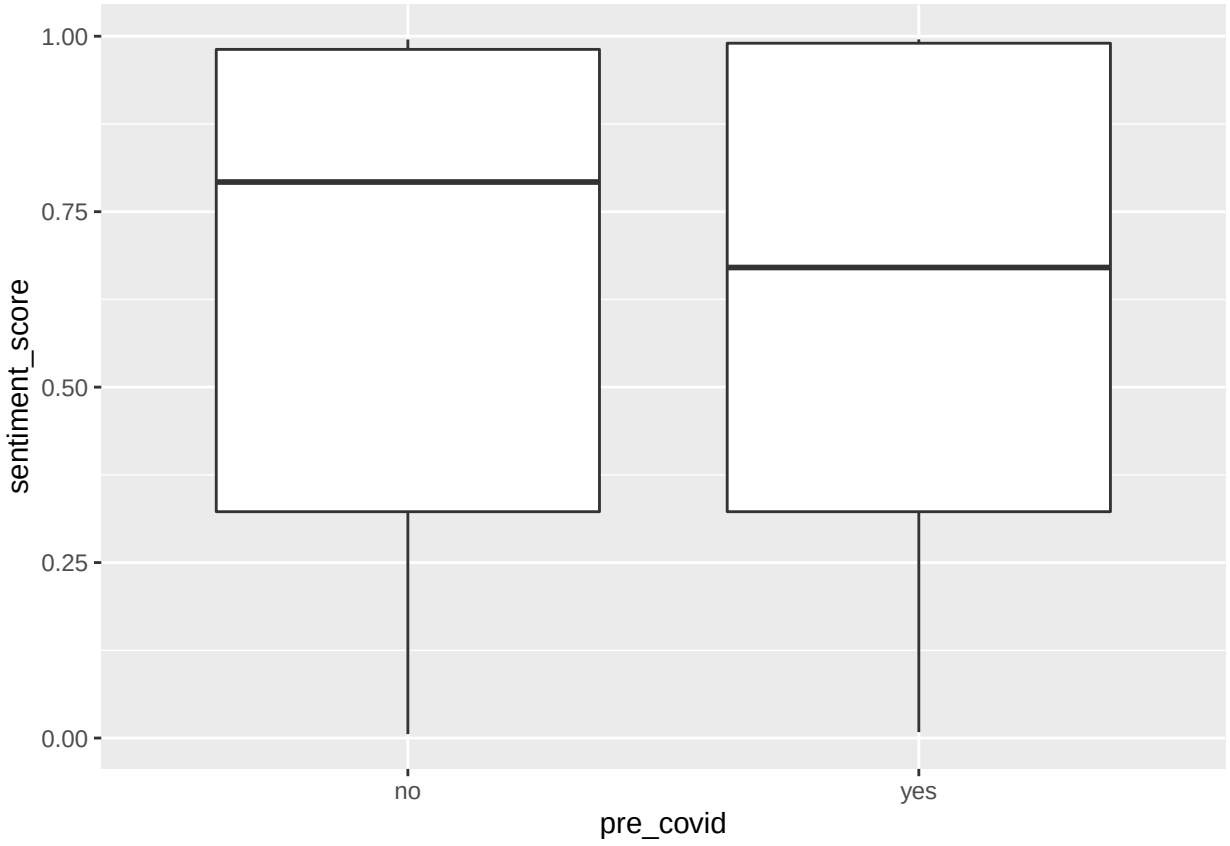

```
ggplot(Canada_analysis_people) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis_people %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.692
##  2         2                   0.606
##  3         3                   0.776
##  4         4                   0.511
##  5         5                   0.410
##  6         6                   0.751
##  7         7                   0.586
##  8         8                   0.755
##  9         9                   0.511
## 10        10                   0.756
## 11        11                   0.610
## 12        12                   0.632
## 13        13                   0.753
```

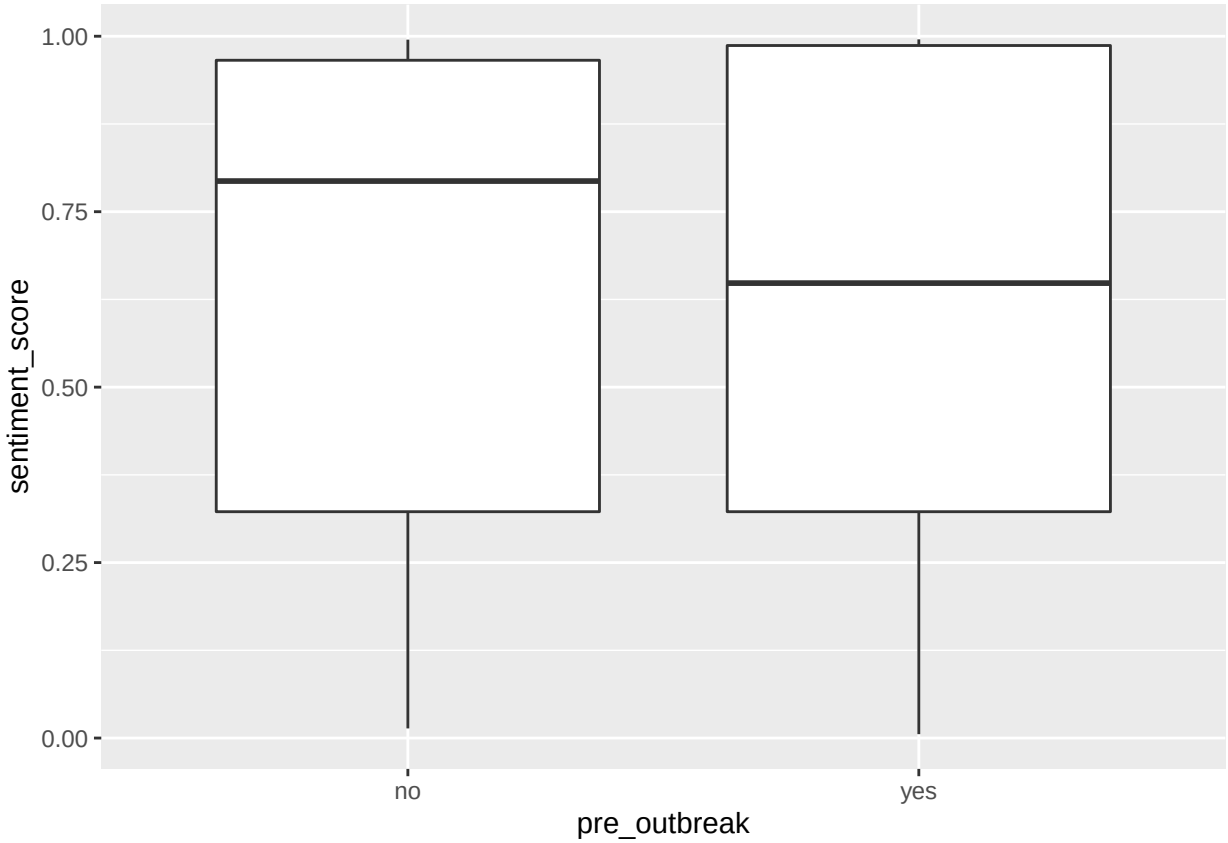
```
ggplot(Canada_analysis_people) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis_people %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.658
## 2 yes                         0.624
```

```
ggplot(Canada_analysis_people) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
Canada_analysis_people %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.666
## 2 yes                            0.635
```

```
#precovid people
count(Canada_analysis_people, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   69
## 2 TRUE                    60
```

```
p_num_precovid = 60
p_num_postcovid = 69
p_num = 129

Canada_analysis_people %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

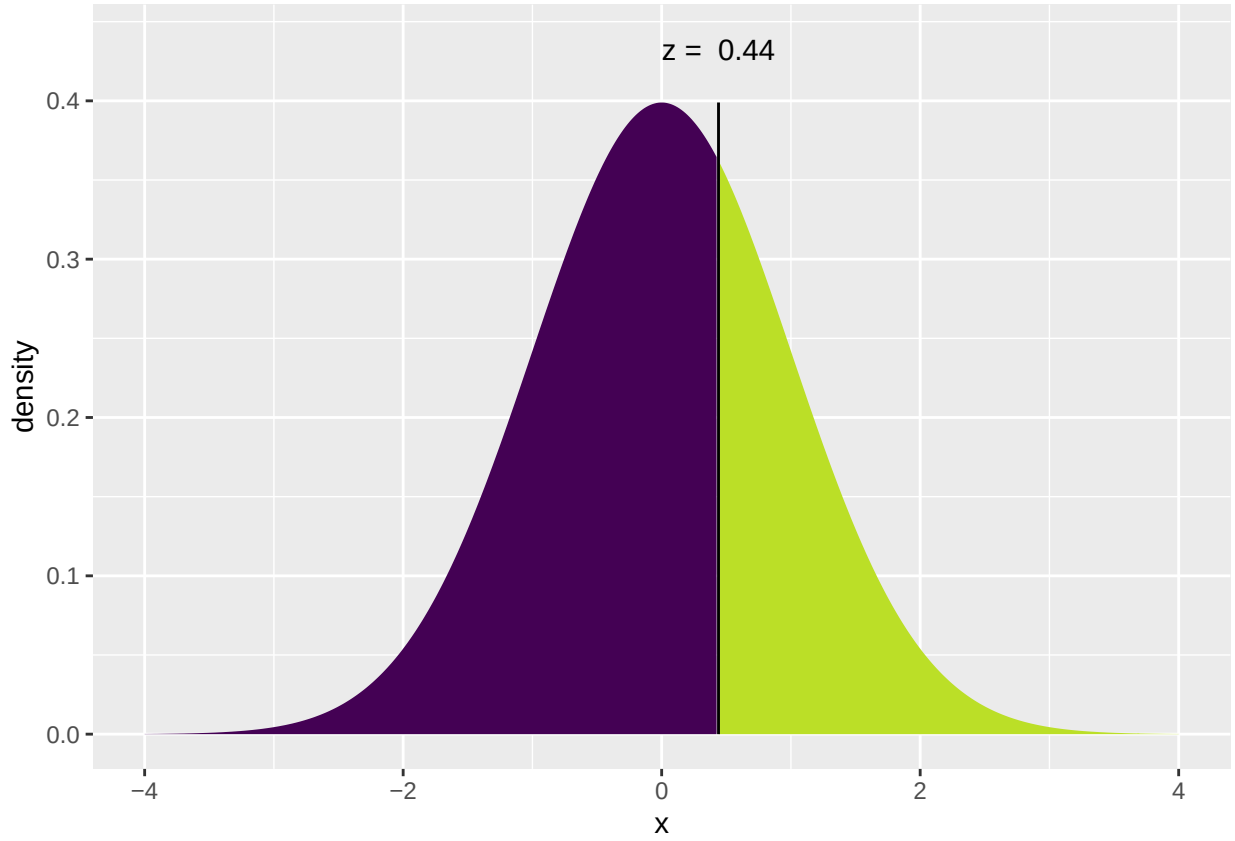
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      24
## 2 TRUE                       36
```

```
p_hat_1_p_pos = 36/60

Canada_analysis_people %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      25
## 2 TRUE                       44
```

```
p_hat_2_p_pos = 44/69

p_hat_p_pos = (36+44)/(60+69)

sd <- sqrt((((p_hat_p_pos)*(1-p_hat_p_pos))/60)+(((p_hat_p_pos)*(1-p_hat_p_pos))/69))
z_score <- ((p_hat_2_p_pos-p_hat_1_p_pos)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.4398) = P(Z <= 0.4398) = 0.67
##  P(X >  0.4398) = P(Z >  0.4398) = 0.33
##
```




```
## [1] 0.6600665
```

```
#outbreak people
count(Canada_analysis_people, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      29
## 2 TRUE                      100
```

```
p_num_preoutbreak = 100
p_num_postoutbreak = 29
p_num = 129

Canada_analysis_people %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      41
## 2 TRUE                       59
```

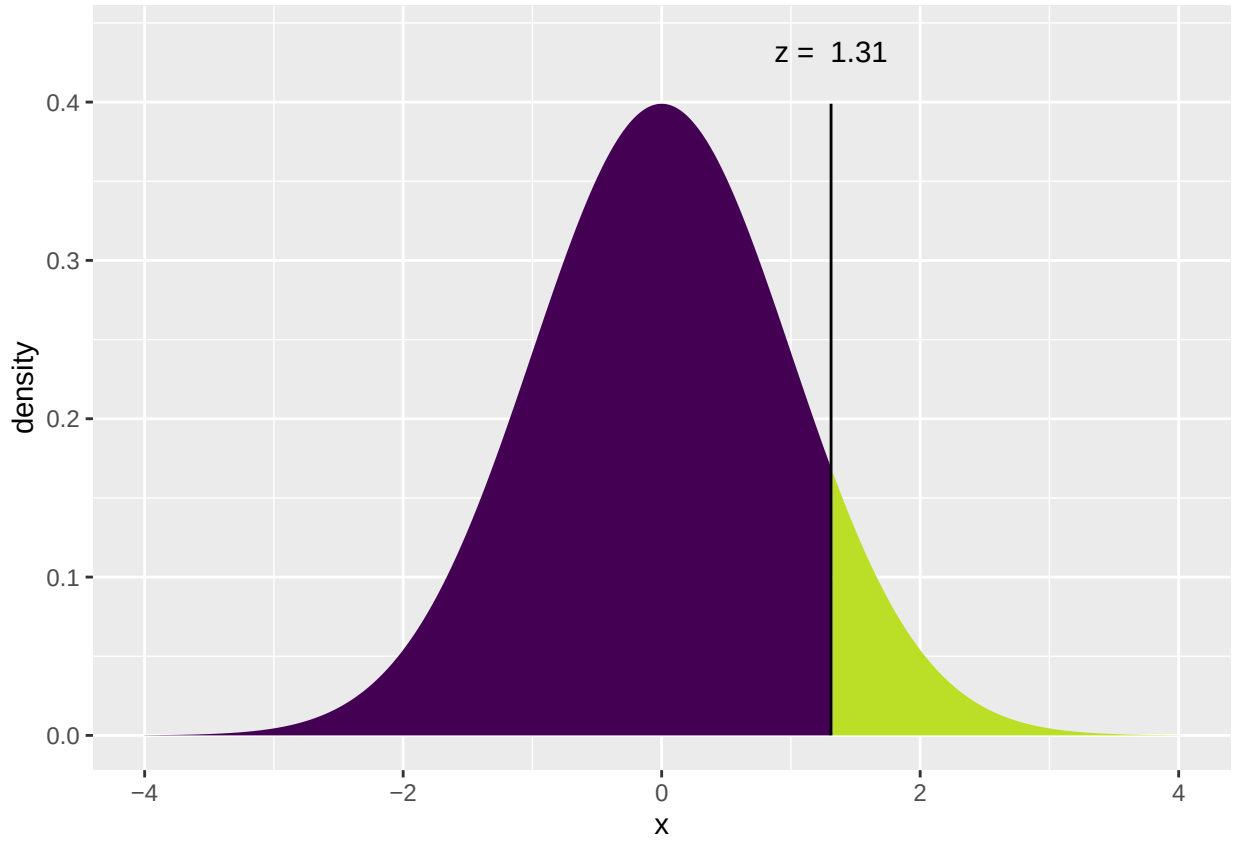
```
p_hat_1_p_pos = 59/100

Canada_analysis_people %>%
```

```
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                       8
## 2 TRUE                       21
```

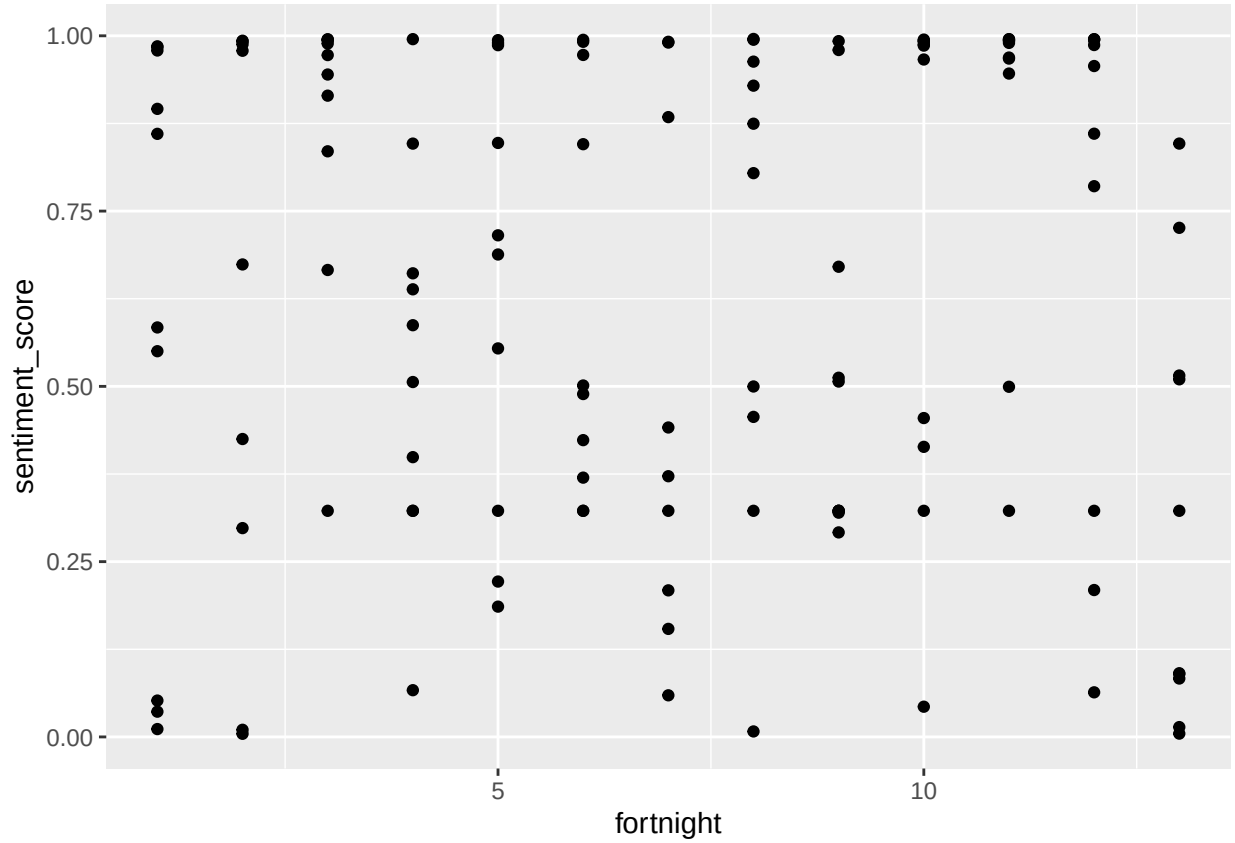
```
p_hat_2_p_pos = 21/29

p_hat_p_pos = (59+21)/(100+29)

sd <- sqrt((((p_hat_p_pos)*(1-p_hat_p_pos))/100)+(((p_hat_p_pos)*(1-p_hat_p_pos))/29))
z_score <- ((p_hat_2_p_pos-p_hat_1_p_pos)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.31) = P(Z <= 1.31) = 0.905
##  P(X >  1.31) = P(Z >  1.31) = 0.09503
##
```

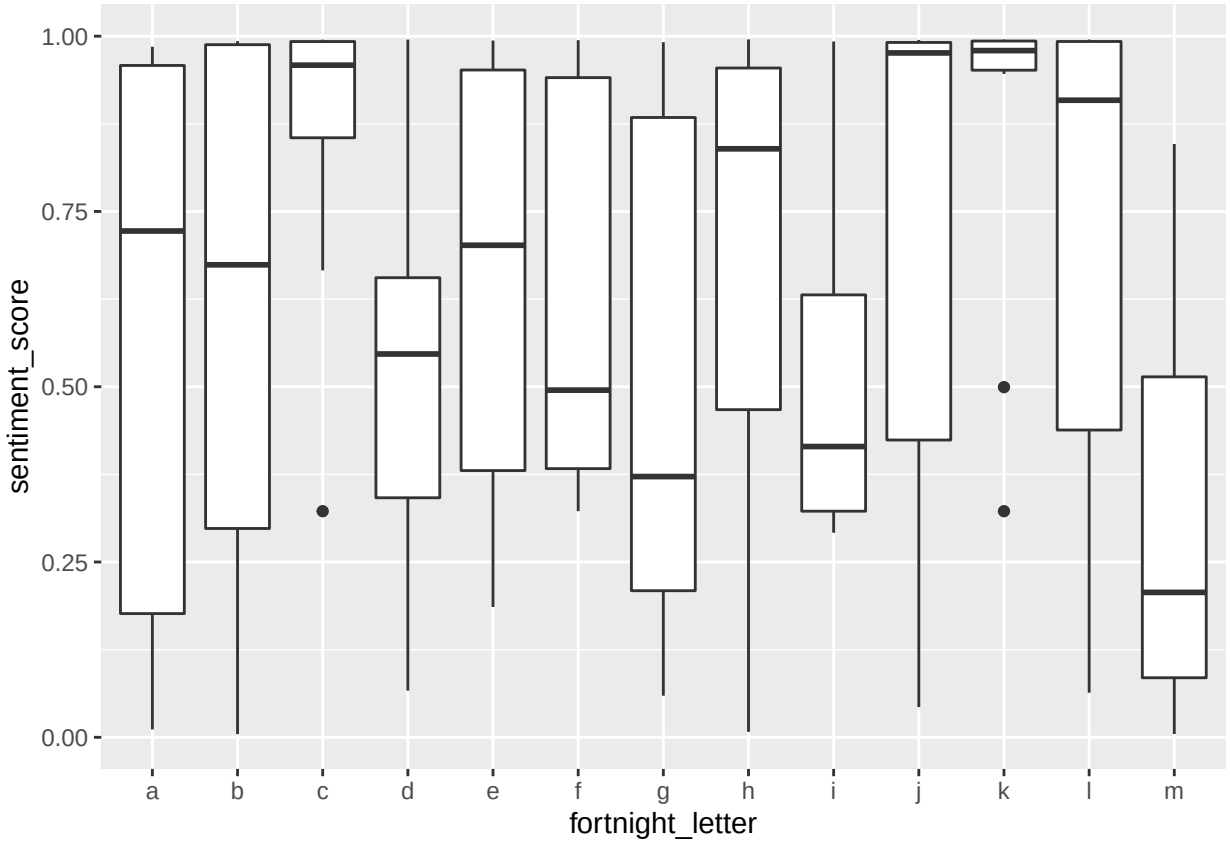


```
## [1] 0.190062
```

```
#data summary entertainment
ggplot(Canada_analysis_entertainment) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



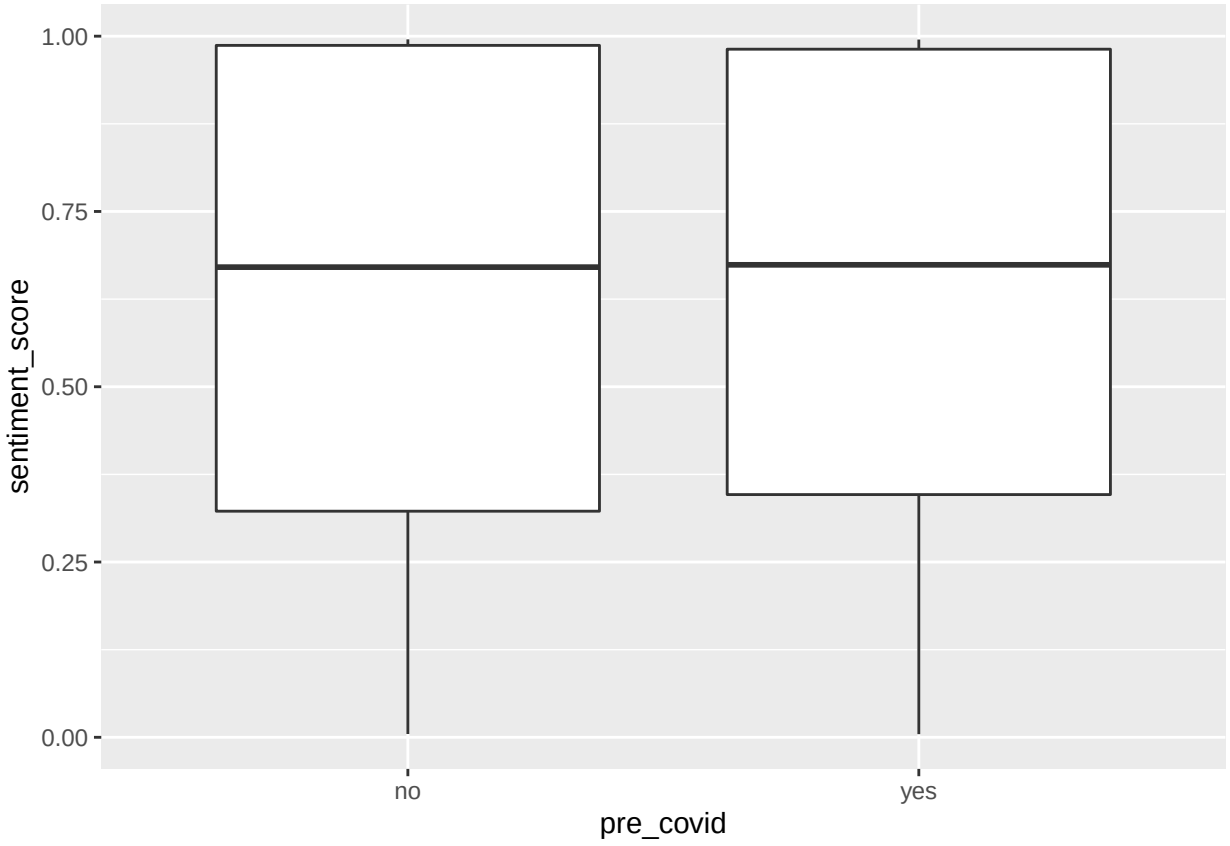

```
ggplot(Canada_analysis_entertainment) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis_entertainment %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.594
##  2         2                   0.596
##  3         3                   0.863
##  4         4                   0.535
##  5         5                   0.651
##  6         6                   0.623
##  7         7                   0.492
##  8         8                   0.685
##  9         9                   0.524
## 10        10                   0.715
## 11        11                   0.867
## 12        12                   0.717
## 13        13                   0.320
```

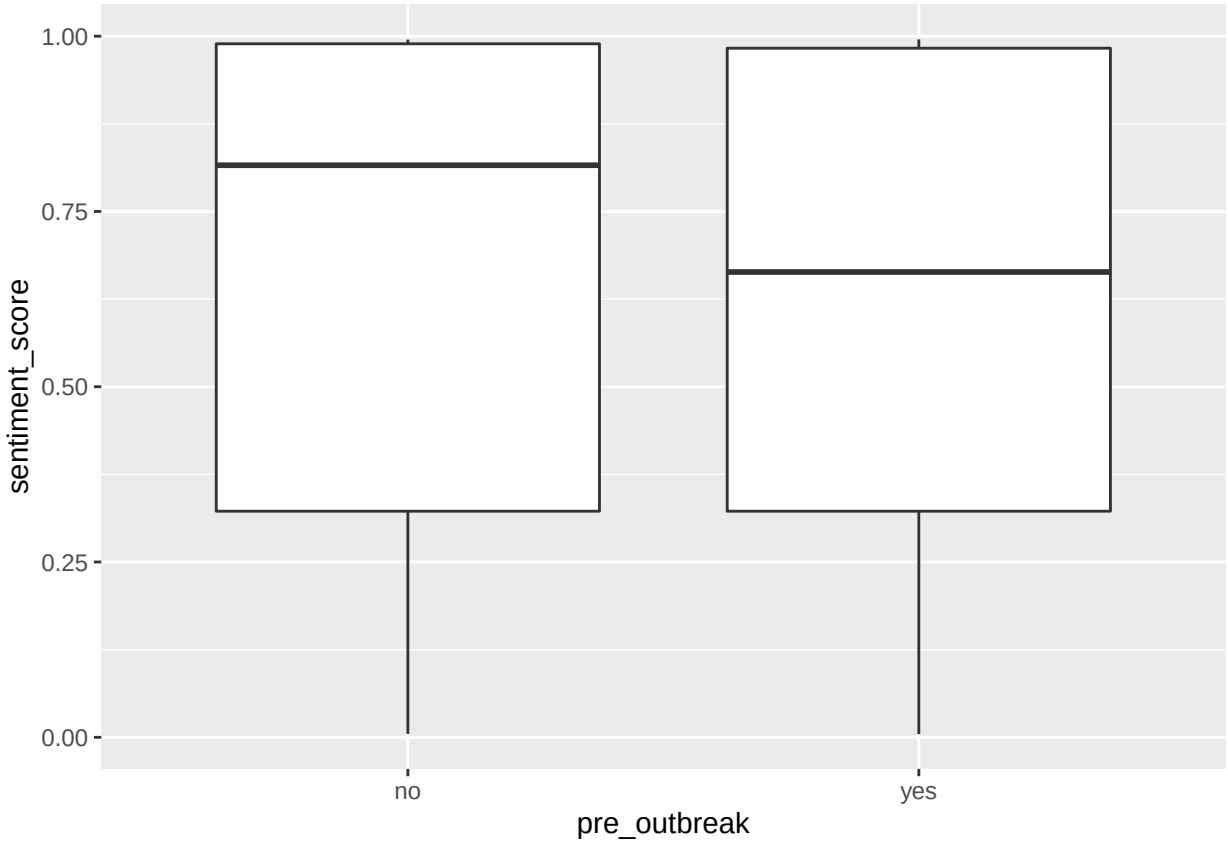
```
ggplot(Canada_analysis_entertainment) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis_entertainment %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.619
## 2 yes                         0.644
```

```
ggplot(Canada_analysis_entertainment) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
Canada_analysis_entertainment %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.635
## 2 yes                            0.629
```

```
#precovid entertainment
count(Canada_analysis_entertainment, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   69
## 2 TRUE                    59
```

```
num_precovid = 59
num_postcovid = 69
num = 128

Canada_analysis_entertainment %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      20
## 2 TRUE                       39
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 39/59

Canada_analysis_entertainment %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       39
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 39/69

p_hat = (39+39)/(59+69)

sd <- sqrt((((p_hat)*(1-p_hat))/59)+(((p_hat)*(1-p_hat))/69))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.107) = P(Z <= -1.107) = 0.1341
##  P(X >  -1.107) = P(Z >  -1.107) = 0.8659
##
```


```
## [1] 0.2681406
```

```
#outbreak entertainment
count(Canada_analysis_entertainment, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       98
```

```
num_preoutbreak = 98
num_postoutbreak = 30
num = 128

Canada_analysis_entertainment %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      39
## 2 TRUE                       59
```

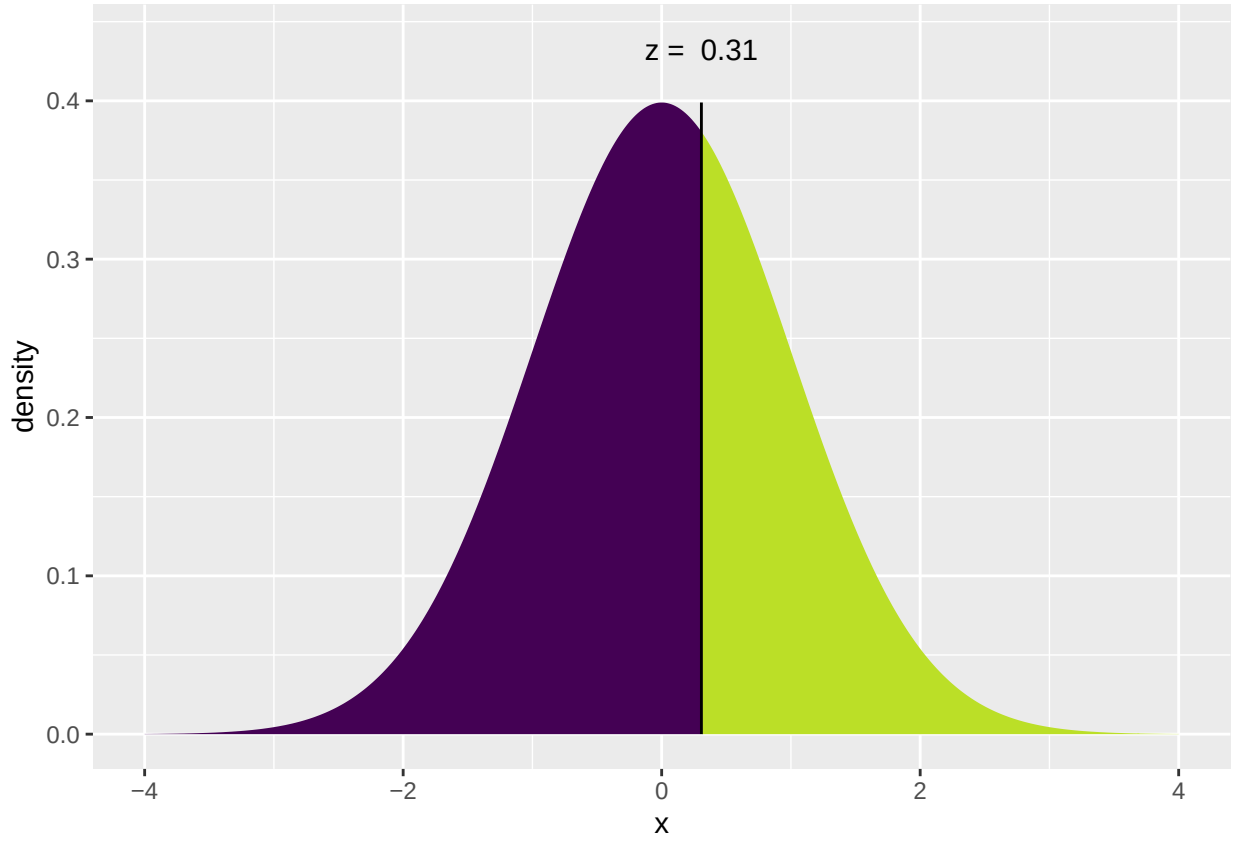
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 59/98
```

```r
Canada_analysis_entertainment %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      11
## 2 TRUE                       19
```

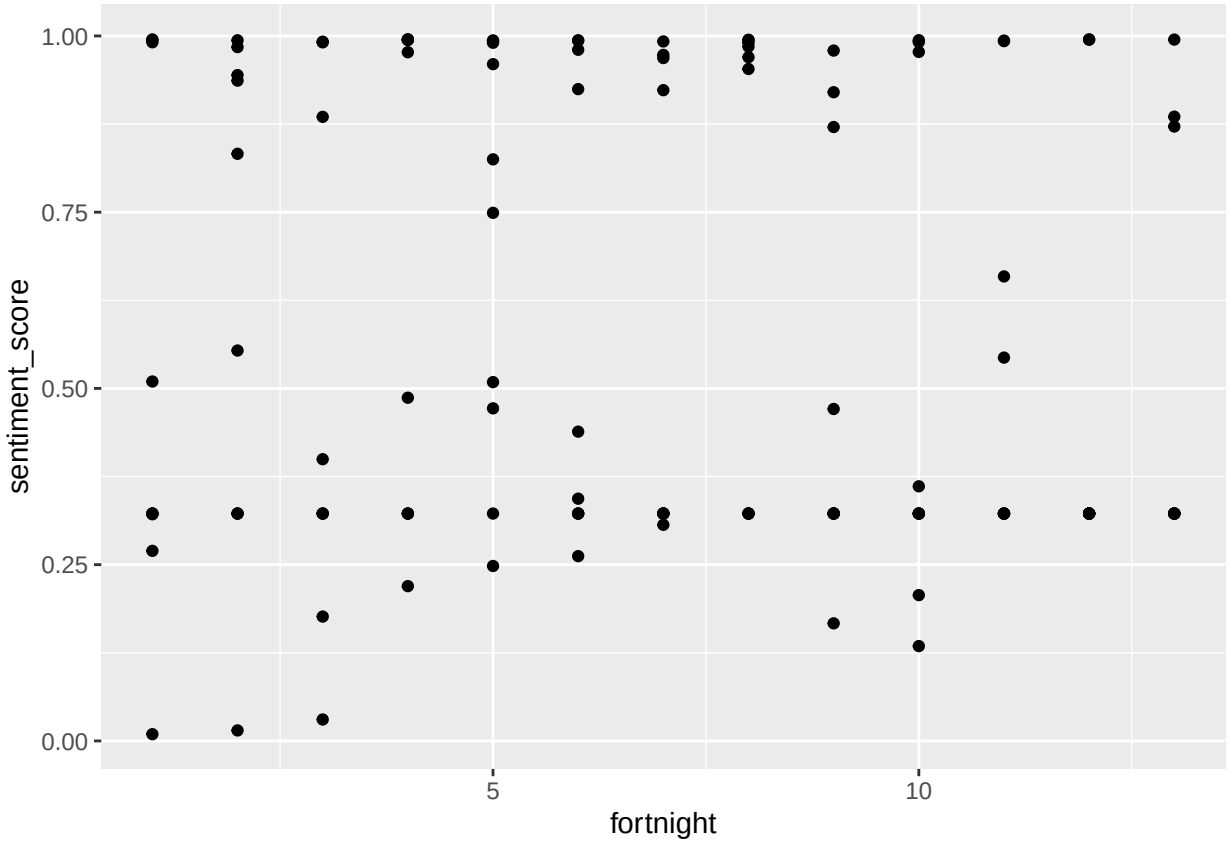
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 19/30

p_hat = (59+19)/(98+30)

sd <- sqrt((((p_hat)*(1-p_hat))/98)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.3074) = P(Z <= 0.3074) = 0.6207
##  P(X >  0.3074) = P(Z >  0.3074) = 0.3793
##
```

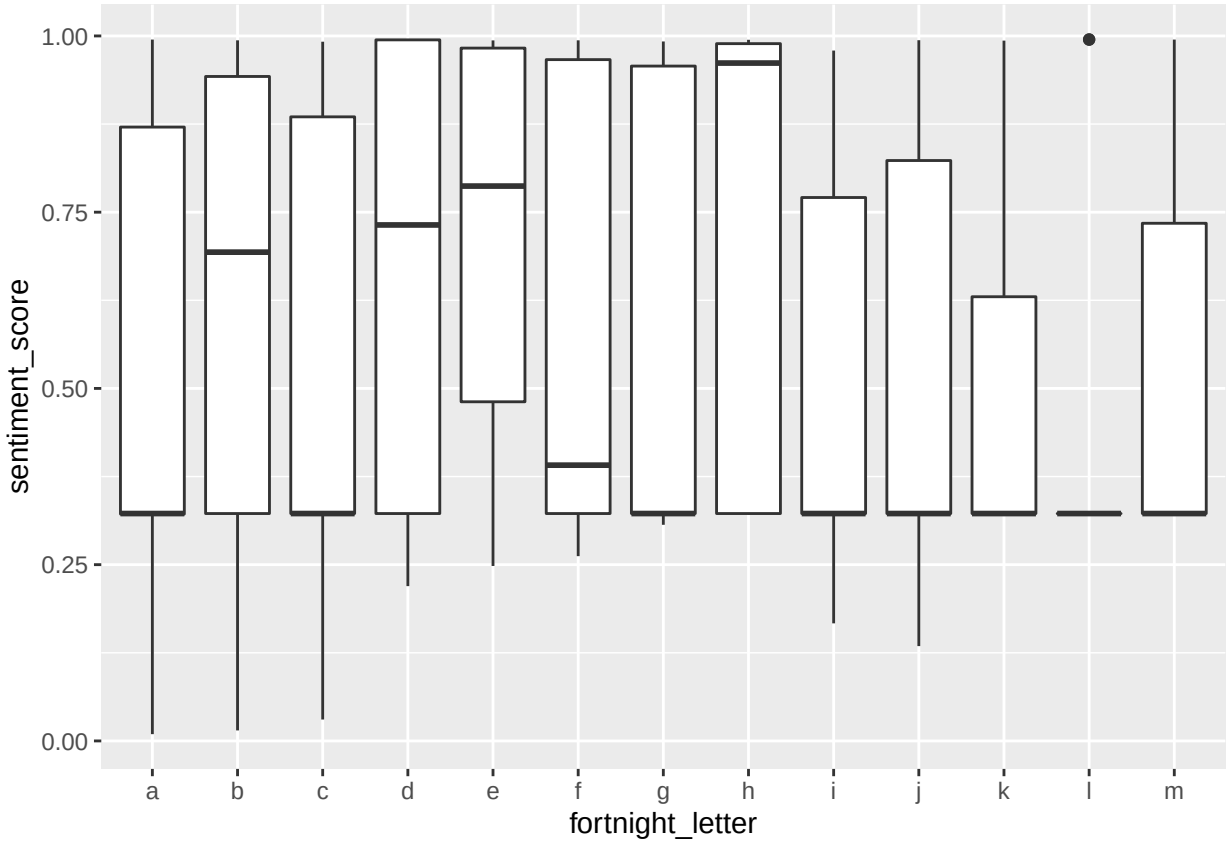



```
## [1] 0.7585481
```

```
#data summary news and politics
ggplot(Canada_analysis_news) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



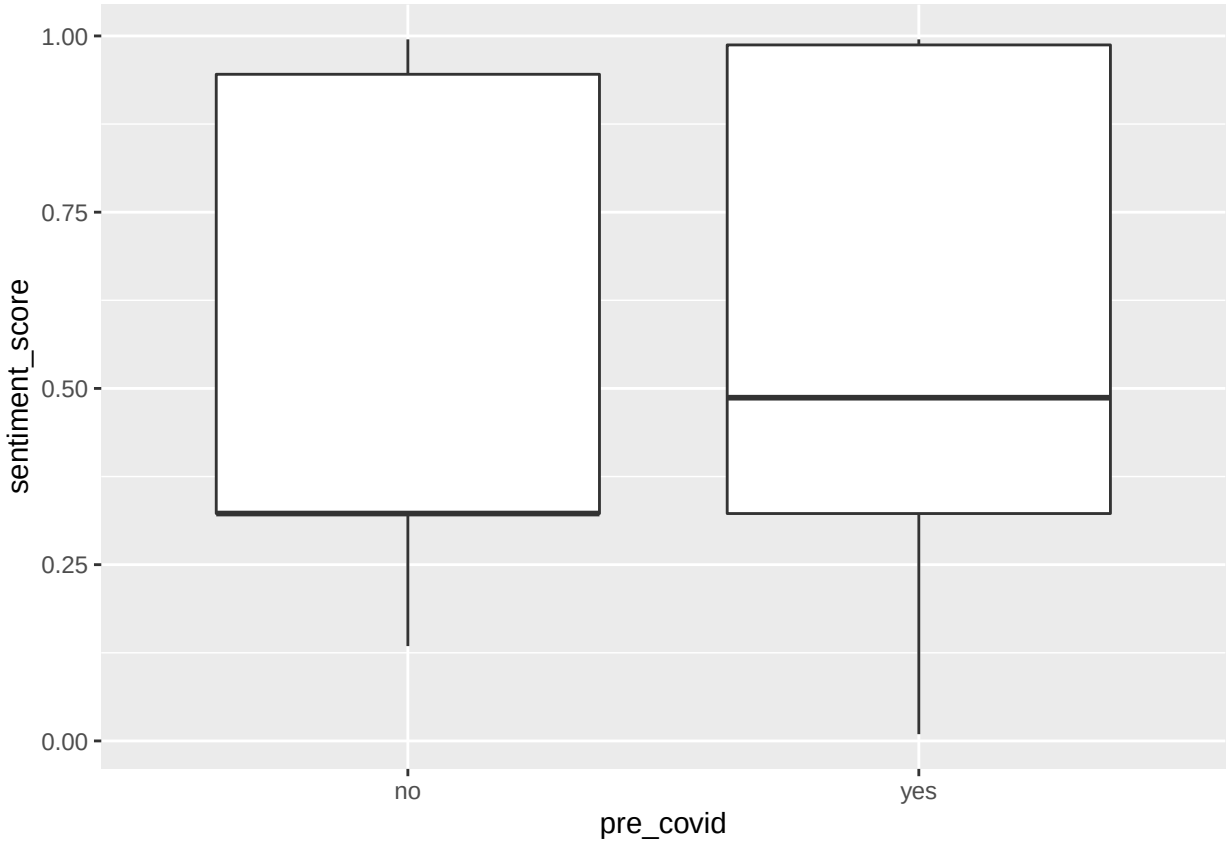

```
ggplot(Canada_analysis_news) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis_news %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.506
##  2         2                   0.623
##  3         3                   0.494
##  4         4                   0.663
##  5         5                   0.706
##  6         6                   0.590
##  7         7                   0.578
##  8         8                   0.718
##  9         9                   0.502
## 10        10                   0.496
## 11        11                   0.512
## 12        12                   0.457
## 13        13                   0.501
```

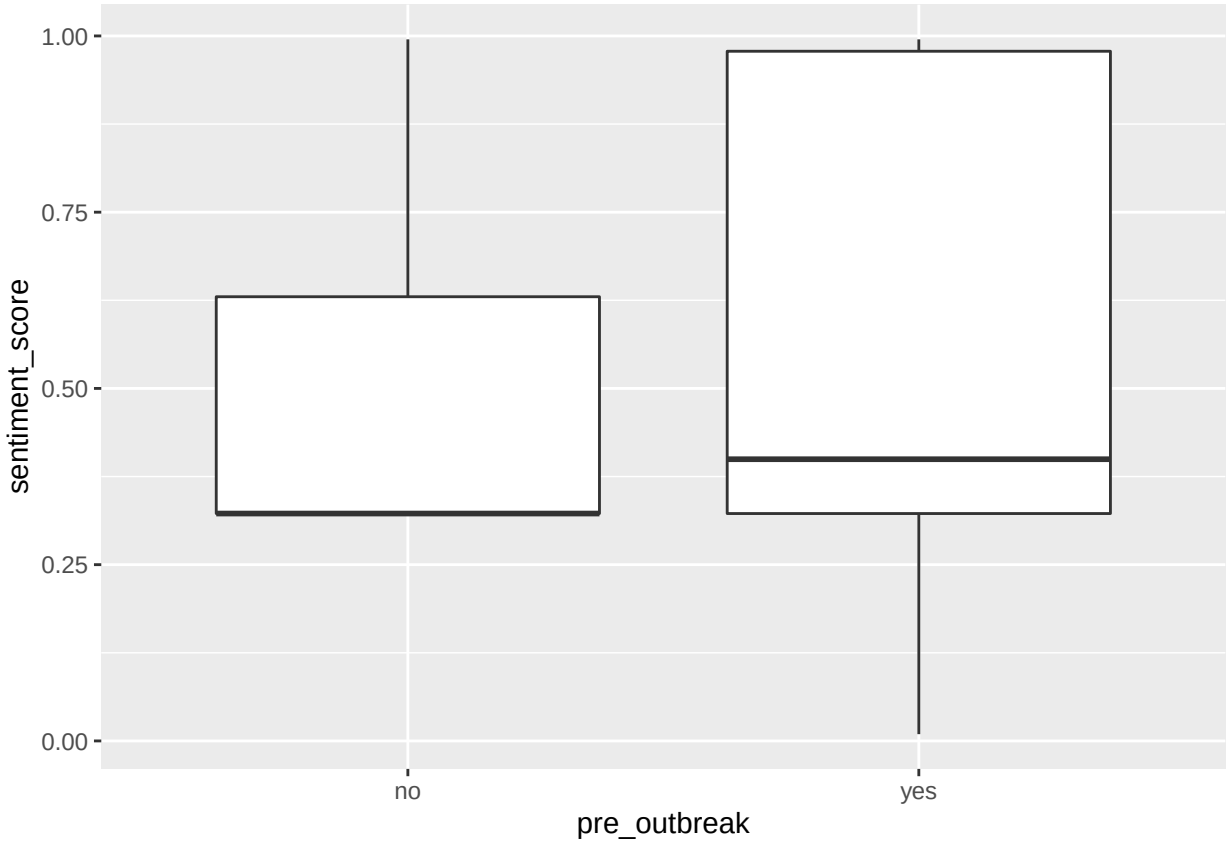
```
ggplot(Canada_analysis_news) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis_news %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.538
## 2 yes                         0.599
```

```
ggplot(Canada_analysis_news) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
Canada_analysis_news %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.490
## 2 yes                            0.588
```

```
#precovid news
count(Canada_analysis_news, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    59
```

```
num_precovid = 59
num_postcovid = 70
num = 129

Canada_analysis_news %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

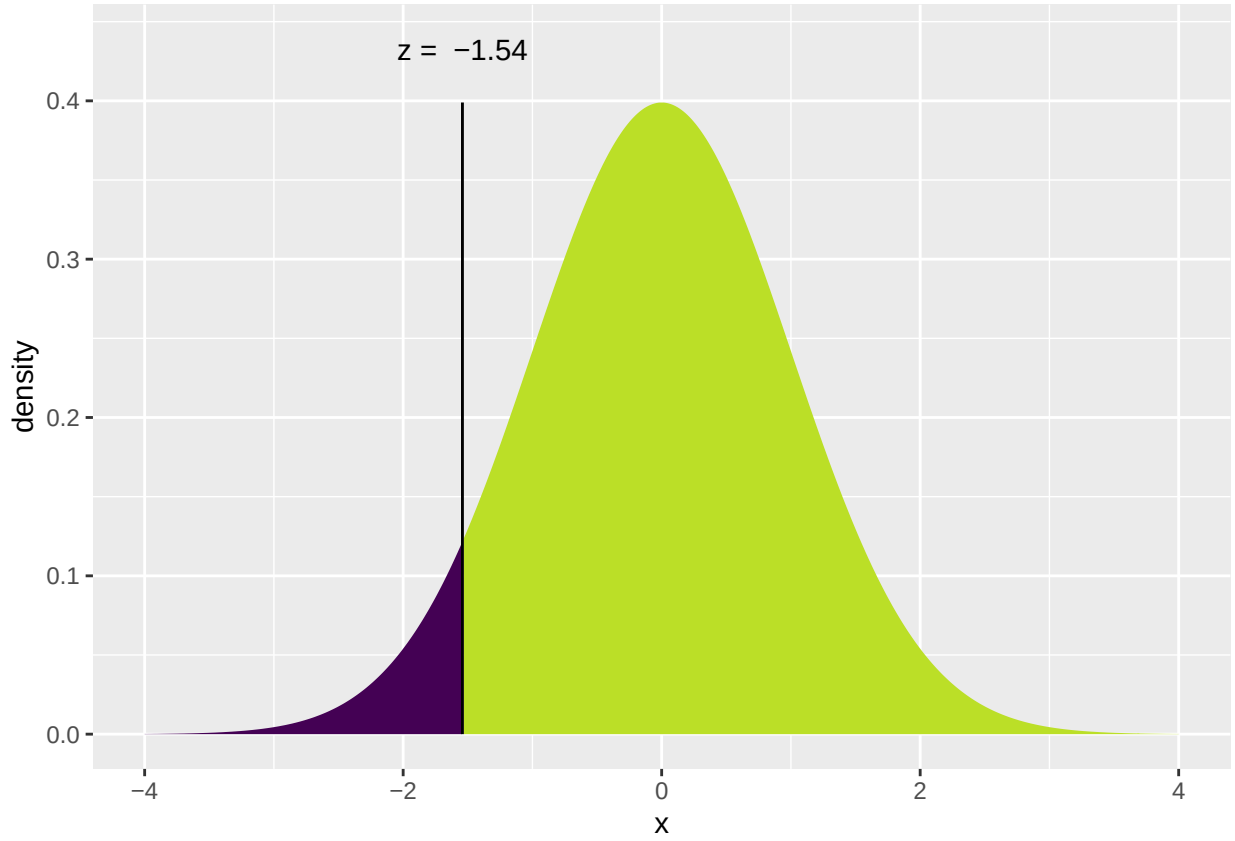
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       29
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 29/59

Canada_analysis_news %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      45
## 2 TRUE                       25
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 25/70

p_hat = (29+25)/(59+70)

sd <- sqrt((((p_hat)*(1-p_hat))/59)+(((p_hat)*(1-p_hat))/70))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.541) = P(Z <= -1.541) = 0.06162
##  P(X >  -1.541) = P(Z >  -1.541) = 0.9384
##
```




```
## [1] 0.1232454
```

```
#outbreak news
count(Canada_analysis_news, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       99
```

```
num_preoutbreak = 99
num_postoutbreak = 30
num = 129

Canada_analysis_news %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      54
## 2 TRUE                       45
```

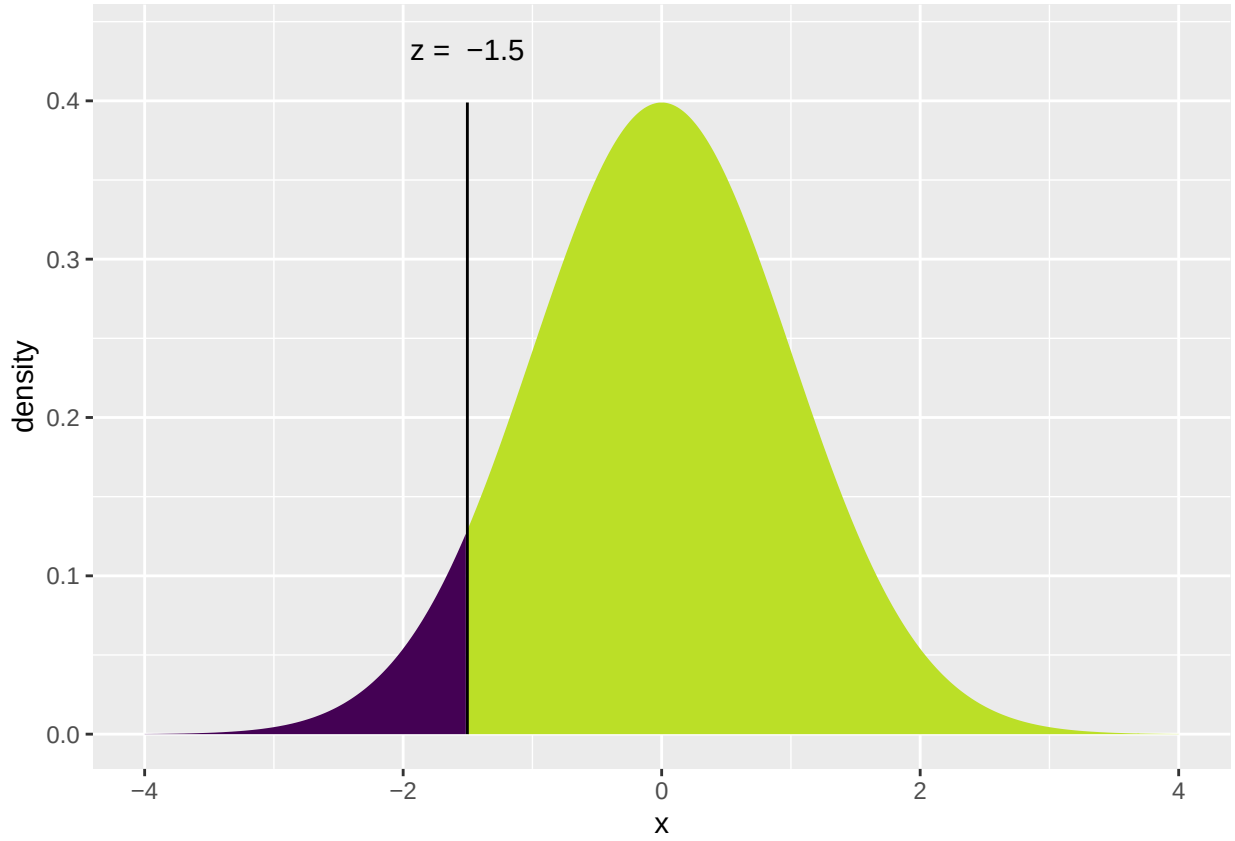
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 45/99
```

```r
Canada_analysis_news %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      21
## 2 TRUE                        9
```

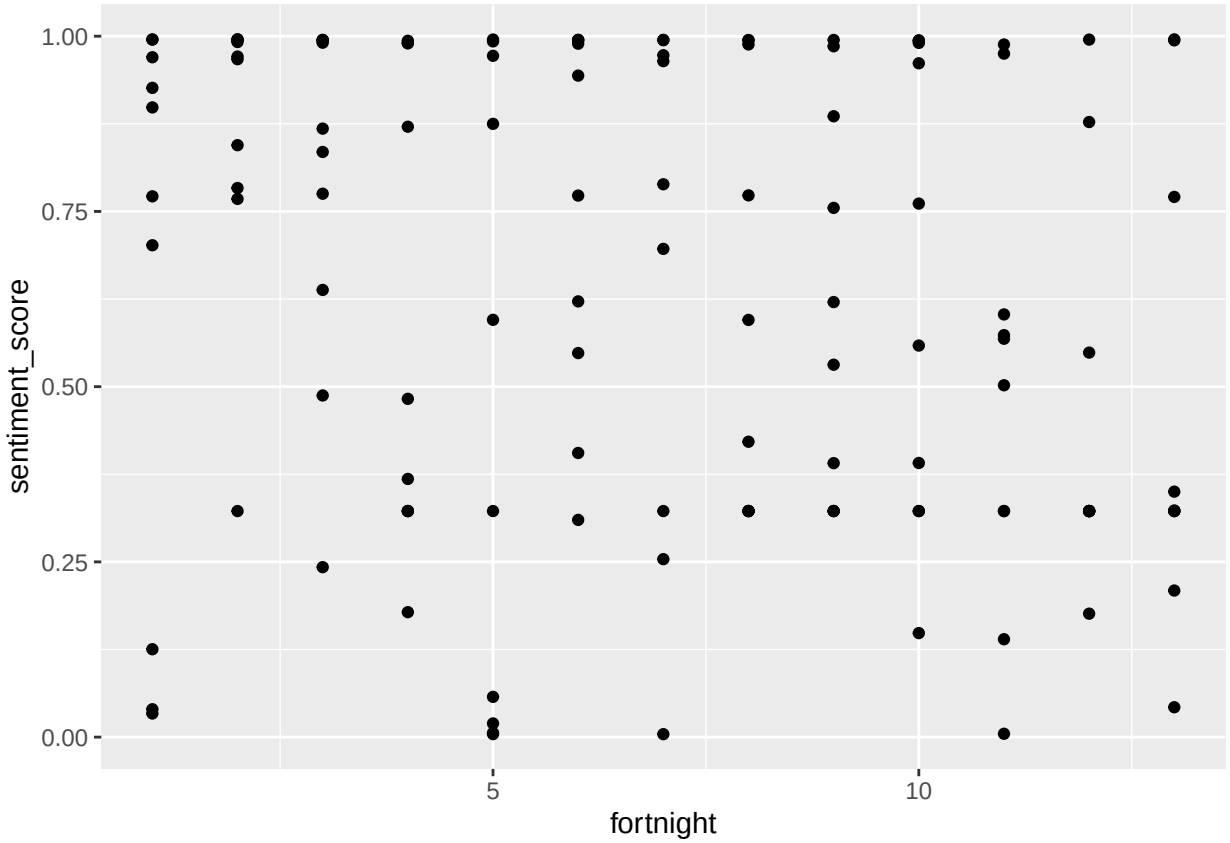
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 9/30

p_hat = (45+9)/(99+30)

sd <- sqrt((((p_hat)*(1-p_hat))/99)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.503) = P(Z <= -1.503) = 0.0664
##  P(X >  -1.503) = P(Z >  -1.503) = 0.9336
##
```

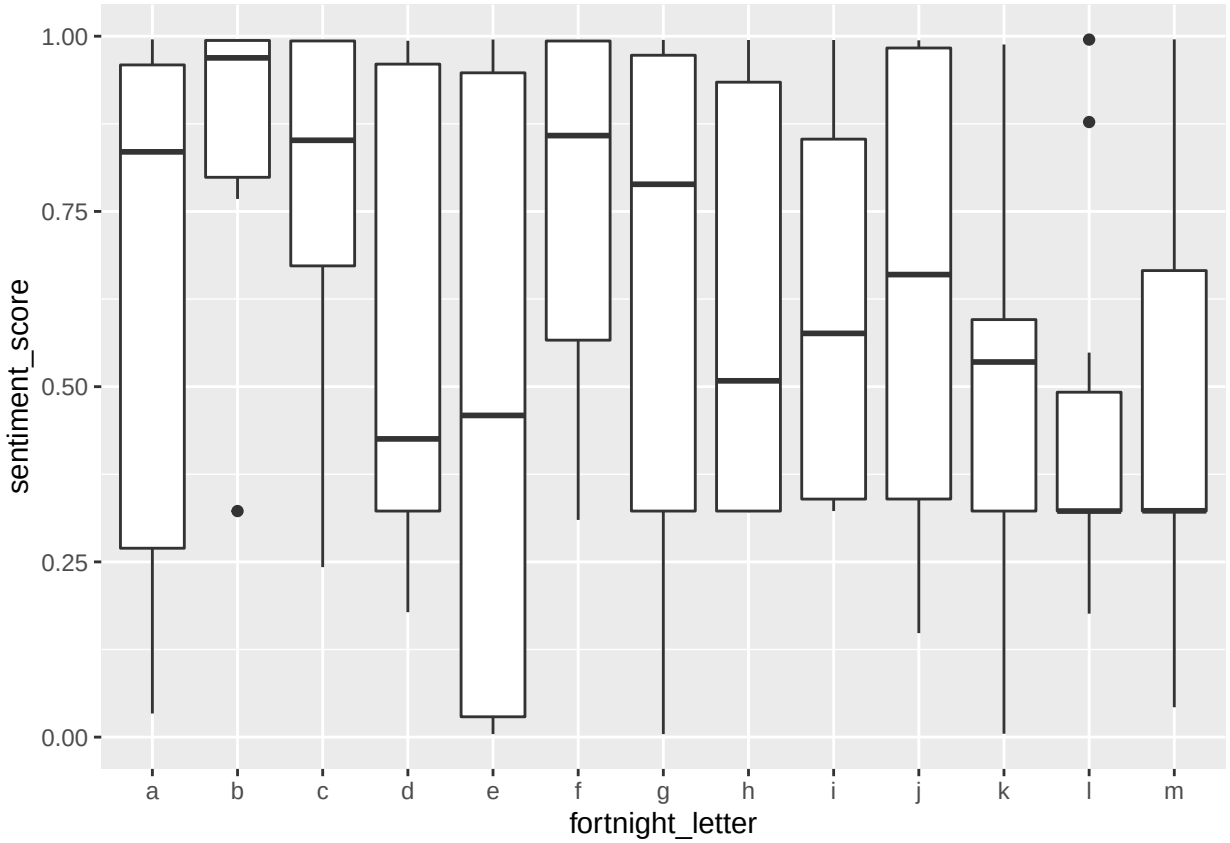



```
## [1] 0.1328008
```

```
#data summary how-to and style
ggplot(Canada_analysis_how_to) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



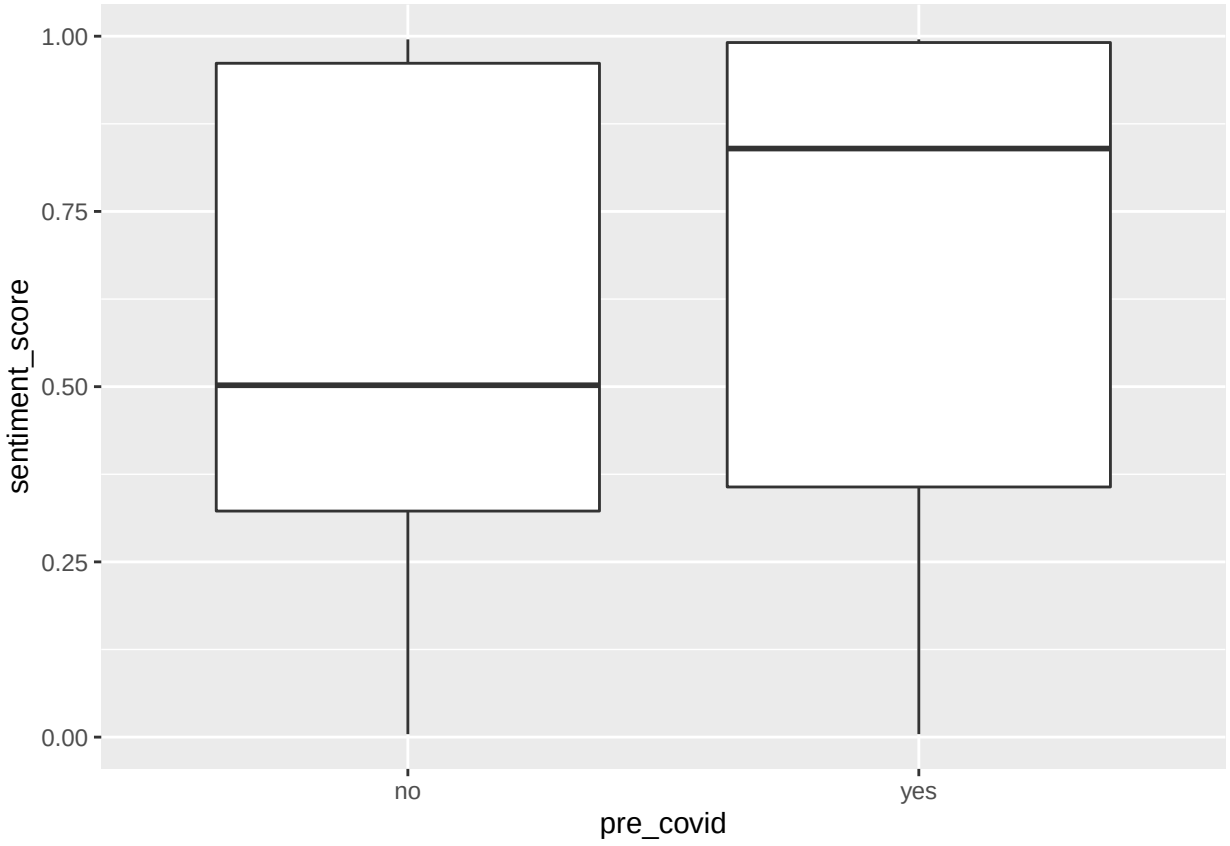

```
ggplot(Canada_analysis_how_to) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis_how_to %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.646
##  2         2                   0.863
##  3         3                   0.782
##  4         4                   0.584
##  5         5                   0.484
##  6         6                   0.757
##  7         7                   0.666
##  8         8                   0.606
##  9         9                   0.613
## 10        10                   0.644
## 11        11                   0.500
## 12        12                   0.453
## 13        13                   0.465
```

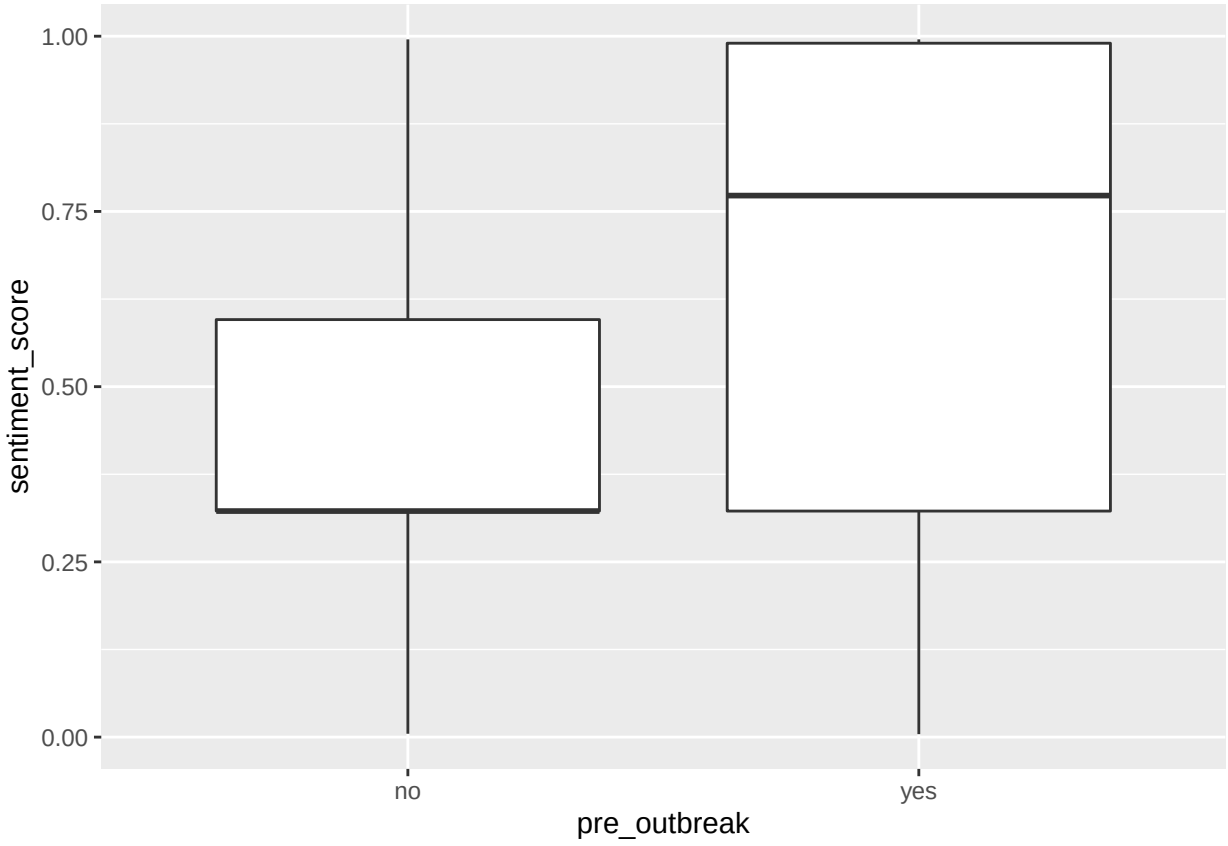
```
ggplot(Canada_analysis_how_to) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis_how_to %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.562
## 2 yes                         0.686
```

```
ggplot(Canada_analysis_how_to) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
Canada_analysis_how_to %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.473
## 2 yes                            0.665
```

```
#precovid how-to
count(Canada_analysis_how_to, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   69
## 2 TRUE                    60
```

```
num_precovid = 60
num_postcovid = 69
num = 129

Canada_analysis_how_to %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

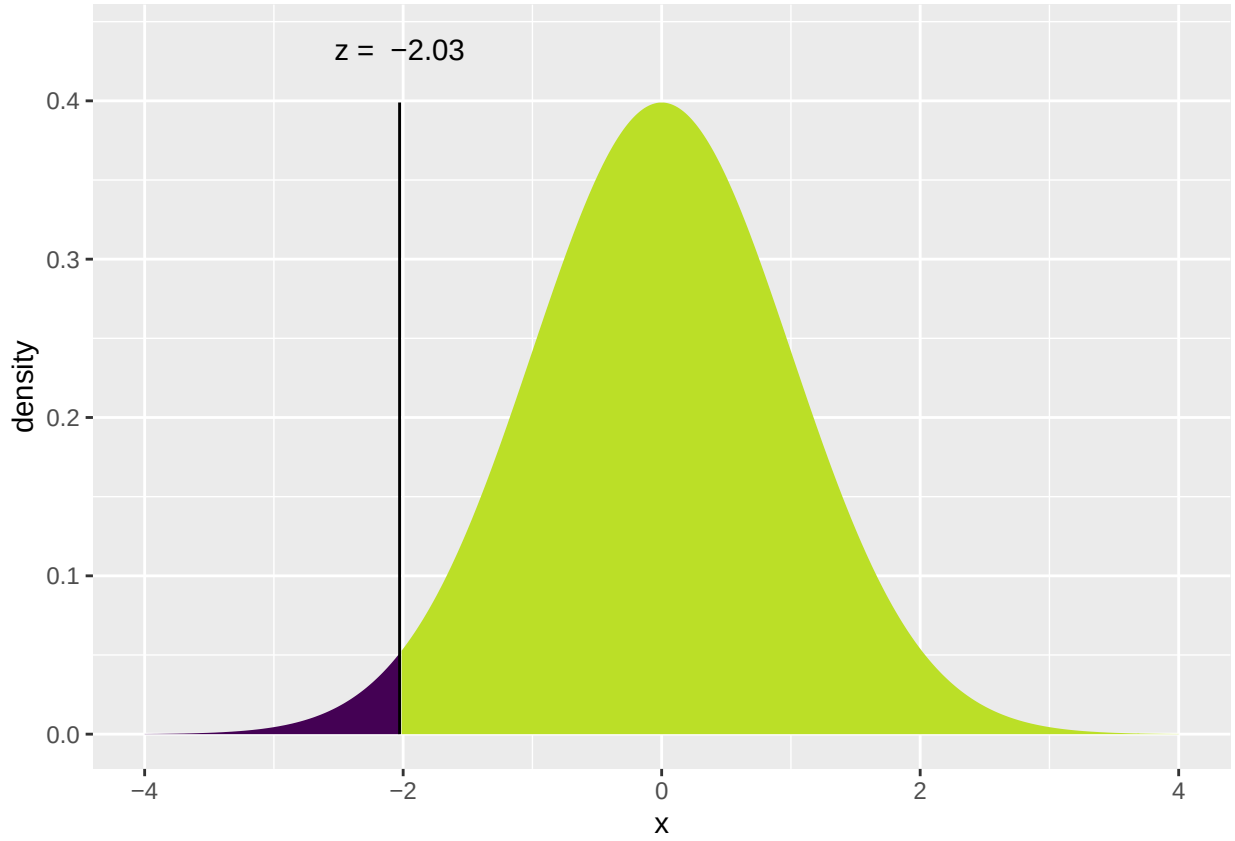
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      19
## 2 TRUE                       41
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 41/60

Canada_analysis_how_to %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      34
## 2 TRUE                       35
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 35/69

p_hat = (41+35)/(60+69)

sd <- sqrt((((p_hat)*(1-p_hat))/60)+(((p_hat)*(1-p_hat))/69))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -2.028) = P(Z <= -2.028) = 0.0213
##  P(X >  -2.028) = P(Z >  -2.028) = 0.9787
##
```




```
## [1] 0.04260338
```

```
#outbreak how-to
count(Canada_analysis_how_to, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       99
```

```
num_preoutbreak = 99
num_postoutbreak = 30
num = 129

Canada_analysis_how_to %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      35
## 2 TRUE                       64
```

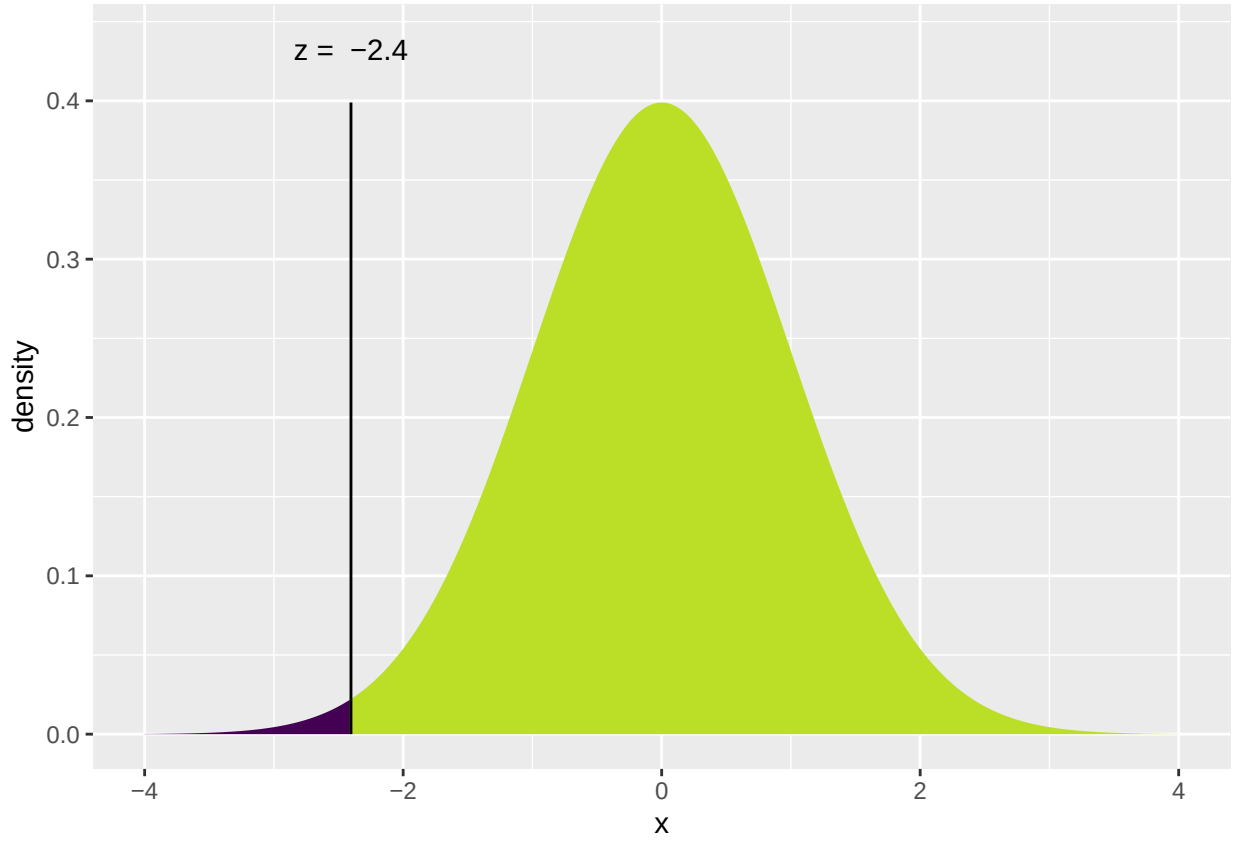
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 64/99
```

```r
Canada_analysis_how_to %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      18
## 2 TRUE                       12
```

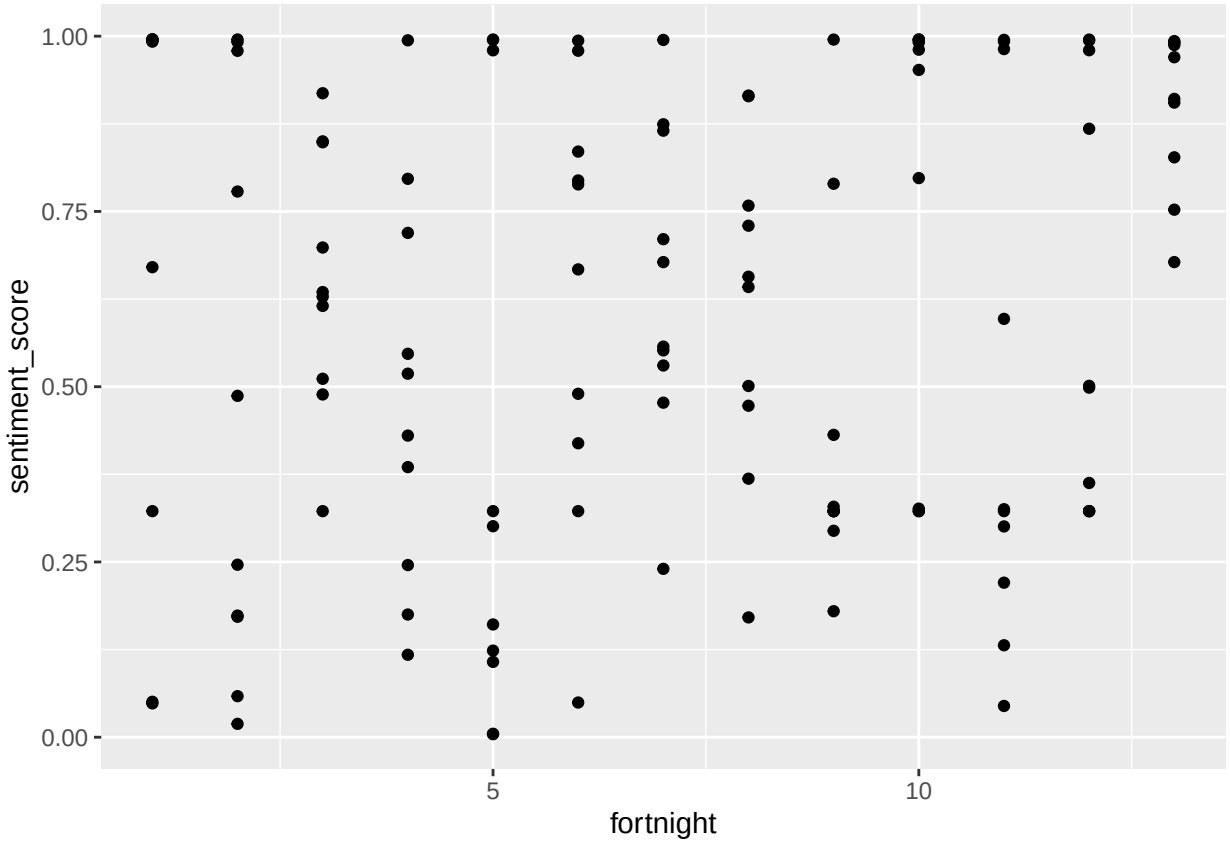
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 12/30

p_hat = (64+12)/(99+30)

sd <- sqrt((((p_hat)*(1-p_hat))/99)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -2.404) = P(Z <= -2.404) = 0.008115
##  P(X >  -2.404) = P(Z >  -2.404) = 0.9919
##
```

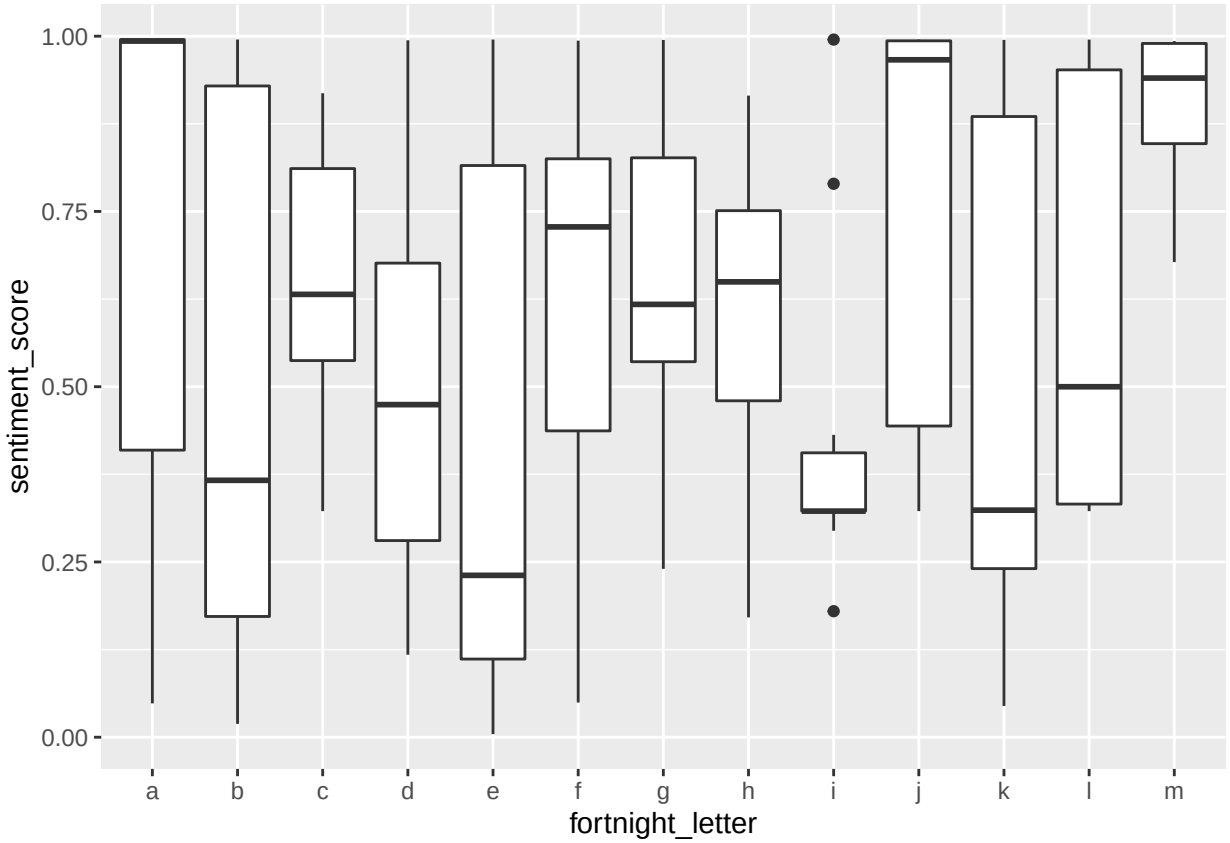



```
## [1] 0.01622942
```

```
#data summary education
ggplot(Canada_analysis_education) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



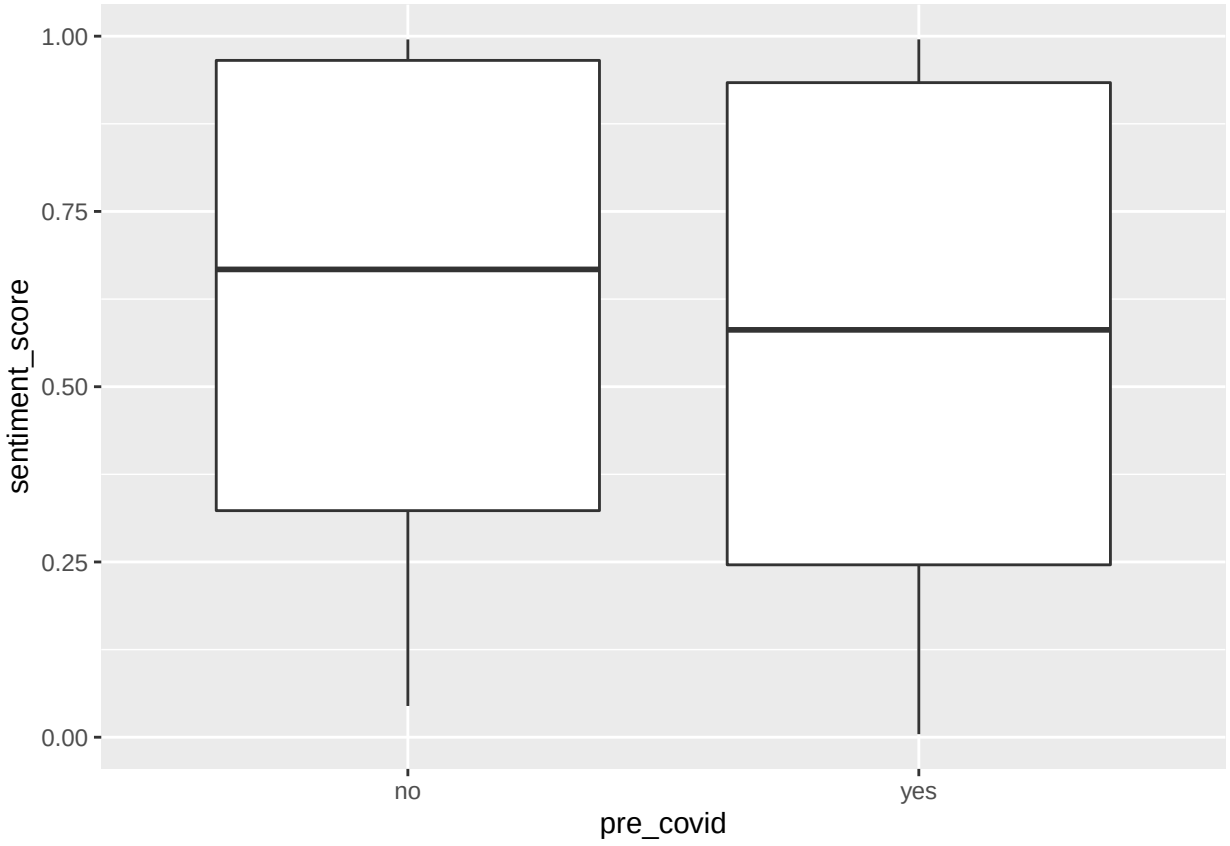

```
ggplot(Canada_analysis_education) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis_education %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.706
##  2         2                   0.490
##  3         3                   0.652
##  4         4                   0.493
##  5         5                   0.399
##  6         6                   0.634
##  7         7                   0.648
##  8         8                   0.613
##  9         9                   0.431
## 10        10                   0.768
## 11        11                   0.491
## 12        12                   0.617
## 13        13                   0.901
```

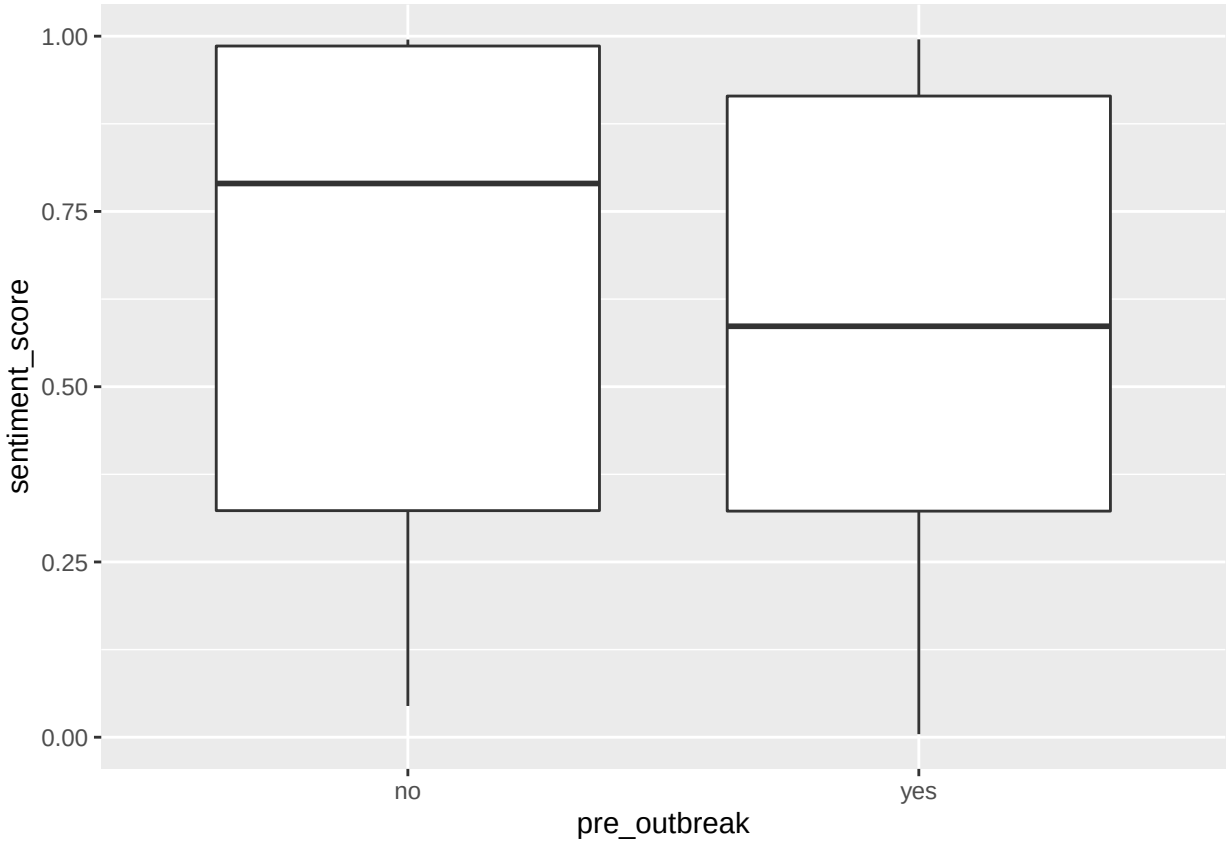
```
ggplot(Canada_analysis_education) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis_education %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.638
## 2 yes                         0.562
```

```
ggplot(Canada_analysis_education) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
Canada_analysis_education %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.669
## 2 yes                            0.583
```

```
#precovid education
count(Canada_analysis_education, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    60
```

```
num_precovid = 60
num_postcovid = 70
num = 130

Canada_analysis_education %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

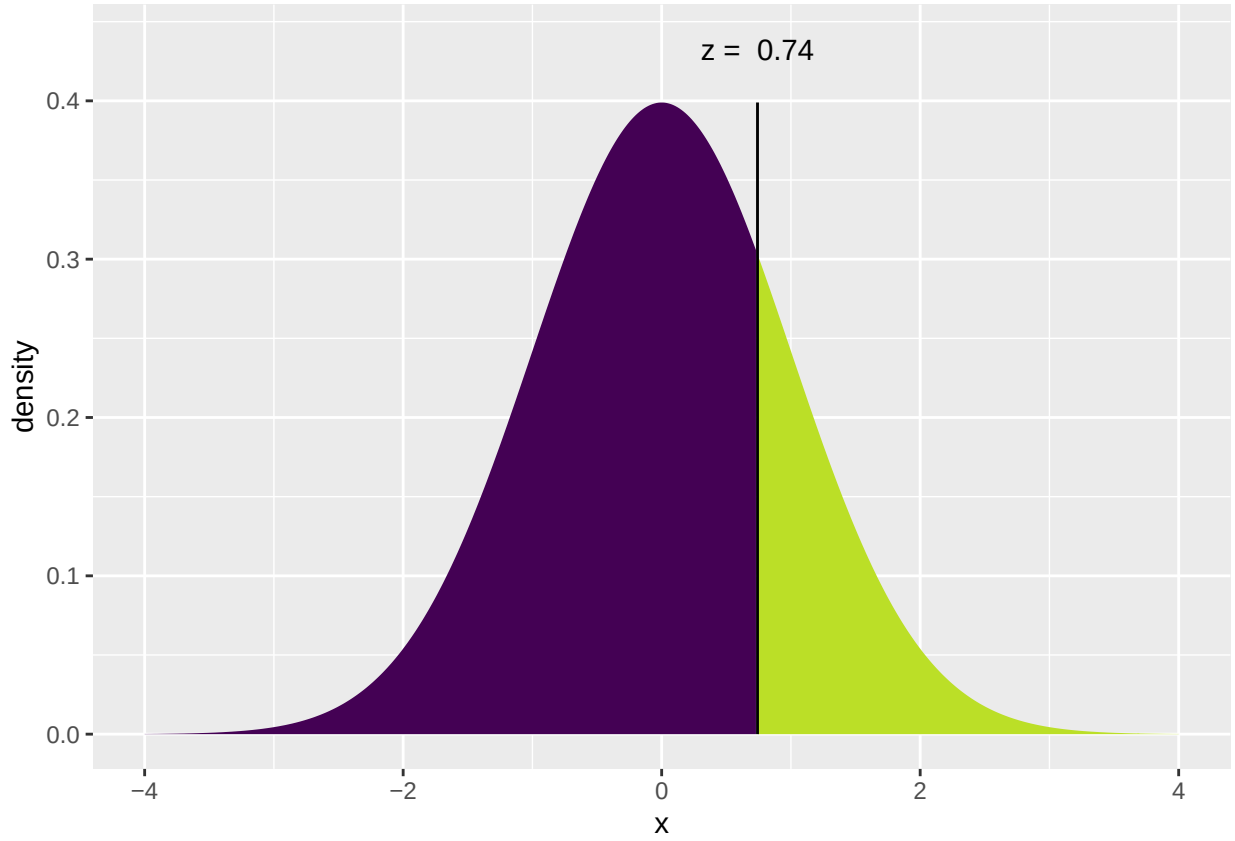
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      27
## 2 TRUE                       33
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 33/60

Canada_analysis_education %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      27
## 2 TRUE                       43
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 43/70

p_hat = (33+43)/(60+70)

sd <- sqrt((((p_hat)*(1-p_hat))/60)+(((p_hat)*(1-p_hat))/70))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.7415) = P(Z <= 0.7415) = 0.7708
##  P(X >  0.7415) = P(Z >  0.7415) = 0.2292
##
```




```
## [1] 0.4583945
```

```
#outbreak education
count(Canada_analysis_education, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                      100
```

```
num_preoutbreak = 100
num_postoutbreak = 30
num = 130

Canada_analysis_education %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      43
## 2 TRUE                       57
```

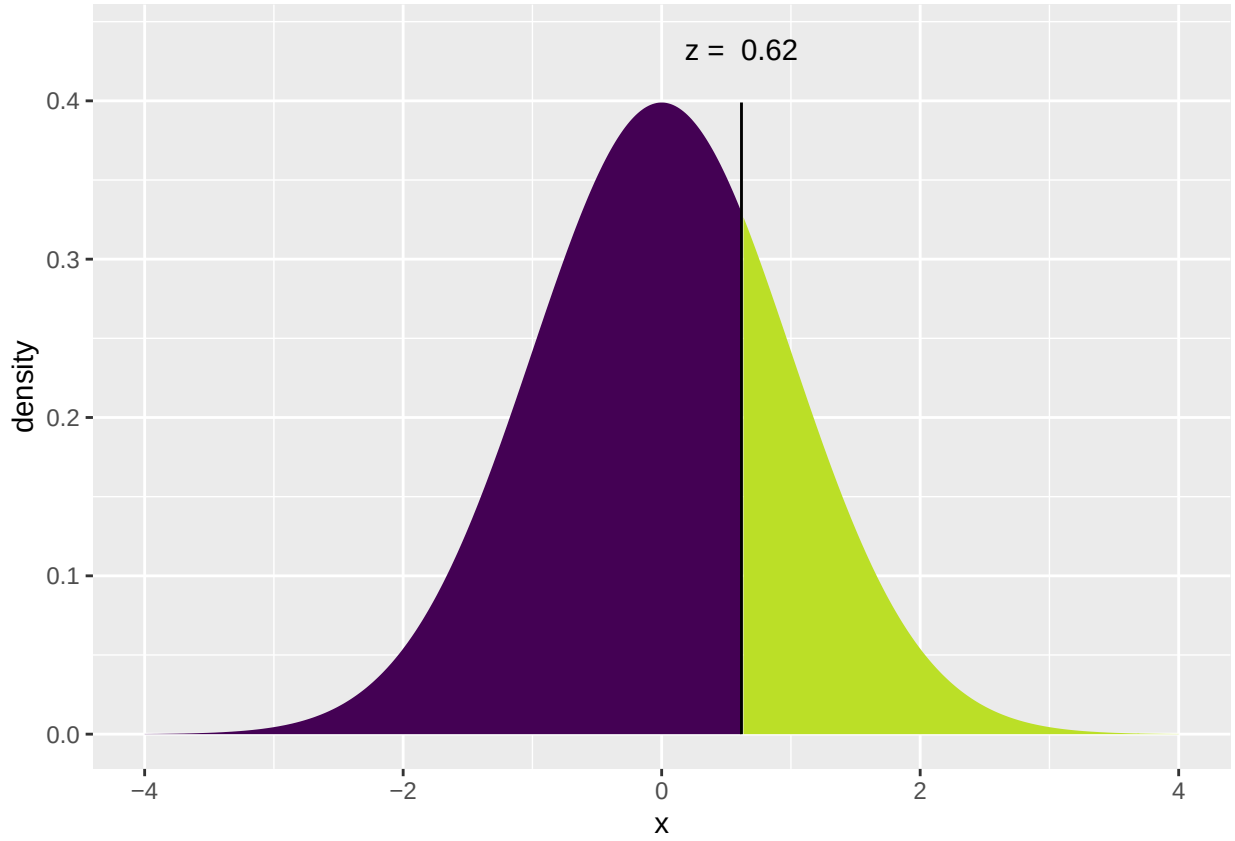
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 57/100
```

```r
Canada_analysis_education %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      11
## 2 TRUE                       19
```

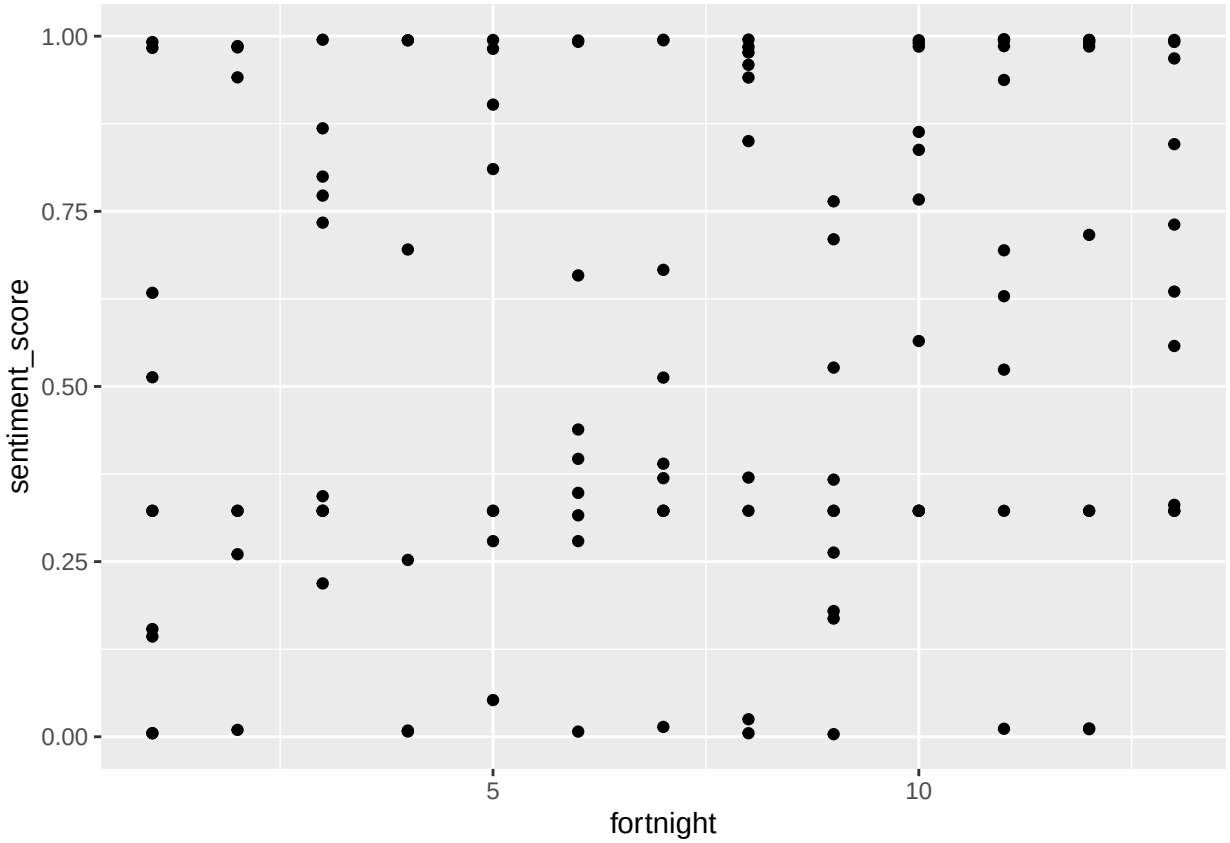
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 19/30

p_hat = (57+19)/(100+30)

sd <- sqrt((((p_hat)*(1-p_hat))/100)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.6174) = P(Z <= 0.6174) = 0.7315
##  P(X >  0.6174) = P(Z >  0.6174) = 0.2685
##
```

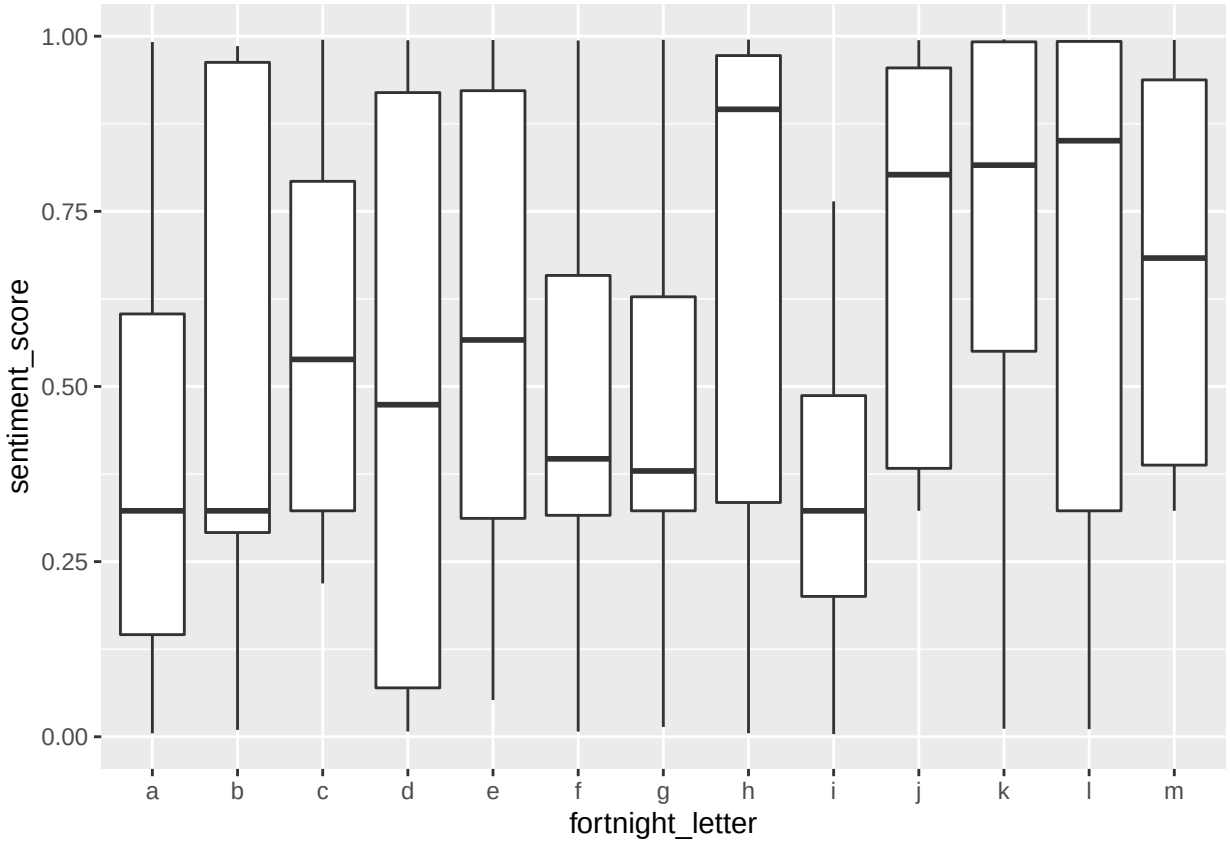



```
## [1] 0.5369762
```

```
#data summary science and technology
ggplot(Canada_analysis_science) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



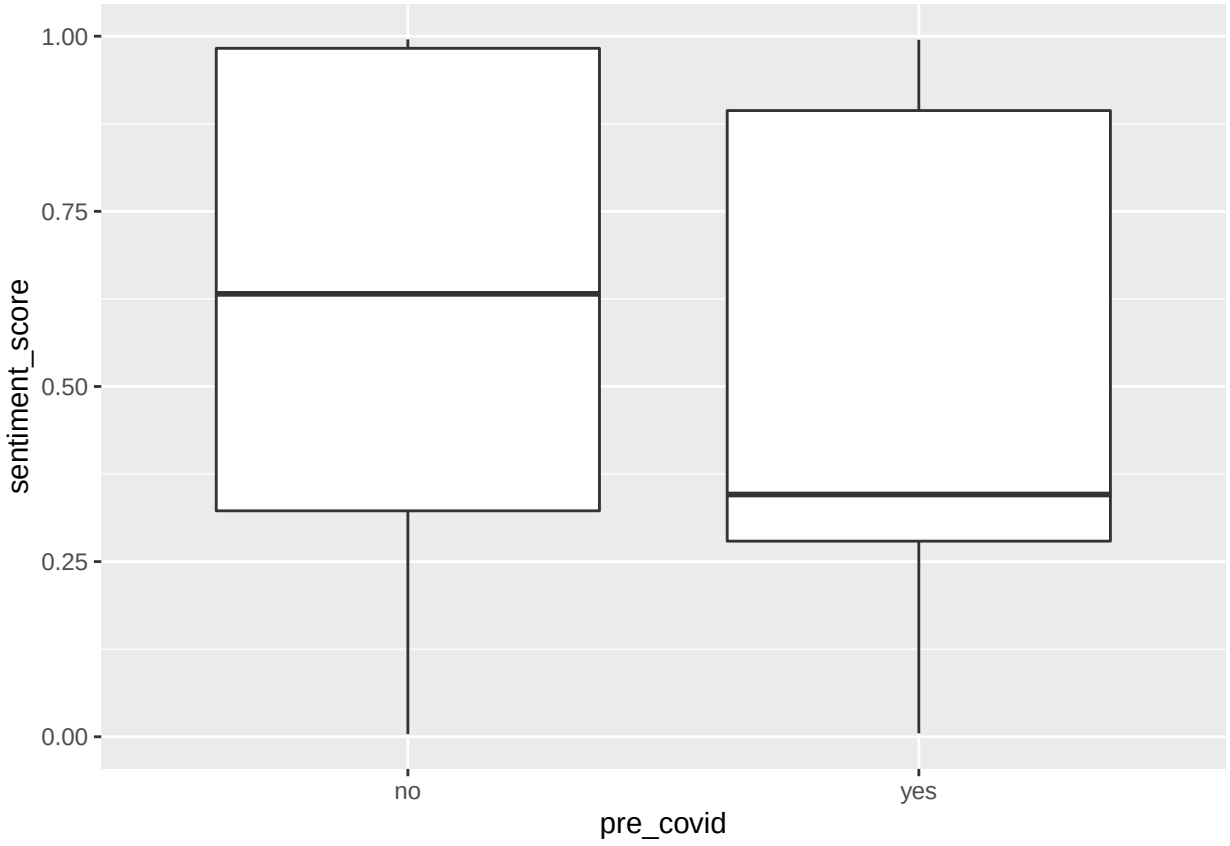

```
ggplot(Canada_analysis_science) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis_science %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.407
##  2         2                   0.547
##  3         3                   0.570
##  4         4                   0.492
##  5         5                   0.583
##  6         6                   0.492
##  7         7                   0.491
##  8         8                   0.643
##  9         9                   0.363
## 10        10                   0.697
## 11        11                   0.709
## 12        12                   0.634
## 13        13                   0.670
```

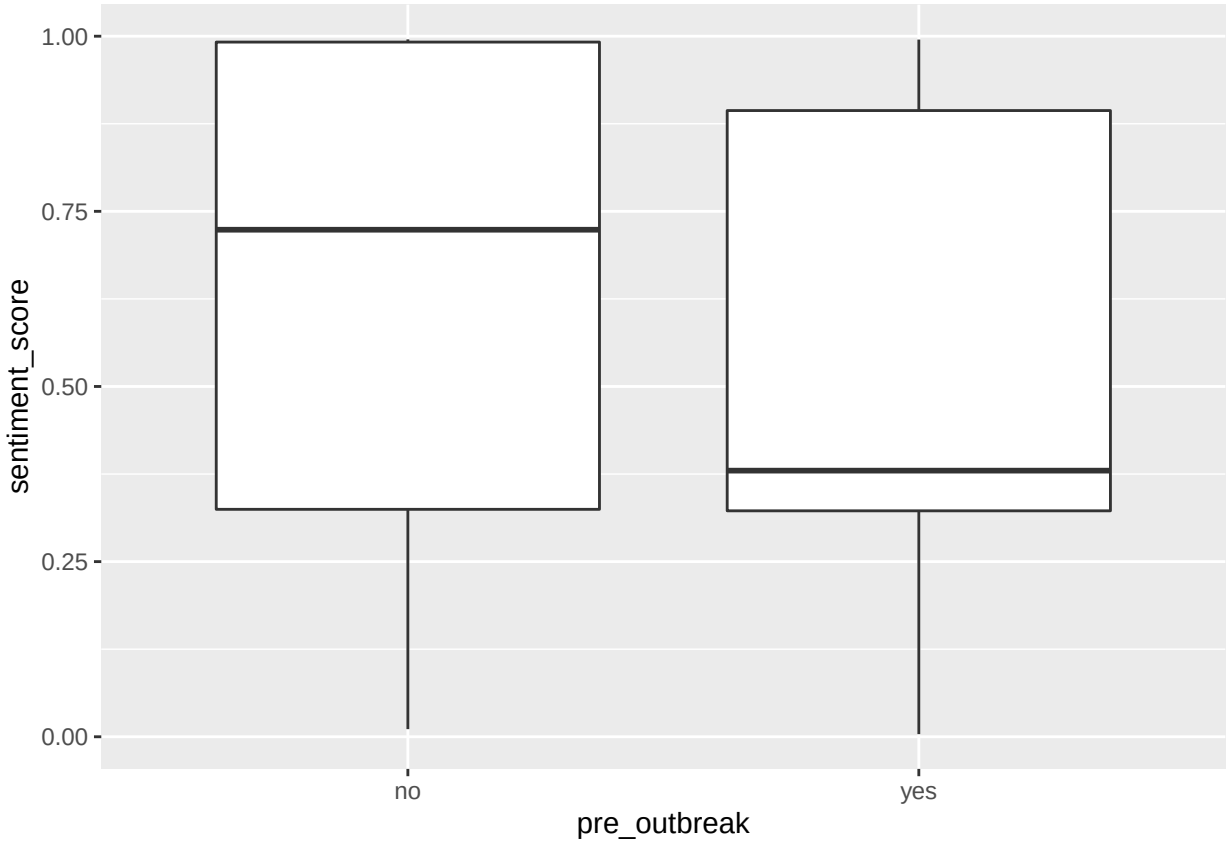
```
ggplot(Canada_analysis_science) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis_science %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.601
## 2 yes                         0.513
```

```
ggplot(Canada_analysis_science) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
Canada_analysis_science %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.671
## 2 yes                            0.529
```

```
#precovid scitech
count(Canada_analysis_science, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   70
## 2 TRUE                    50
```

```
num_precovid = 50
num_postcovid = 70
num = 120

Canada_analysis_science %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

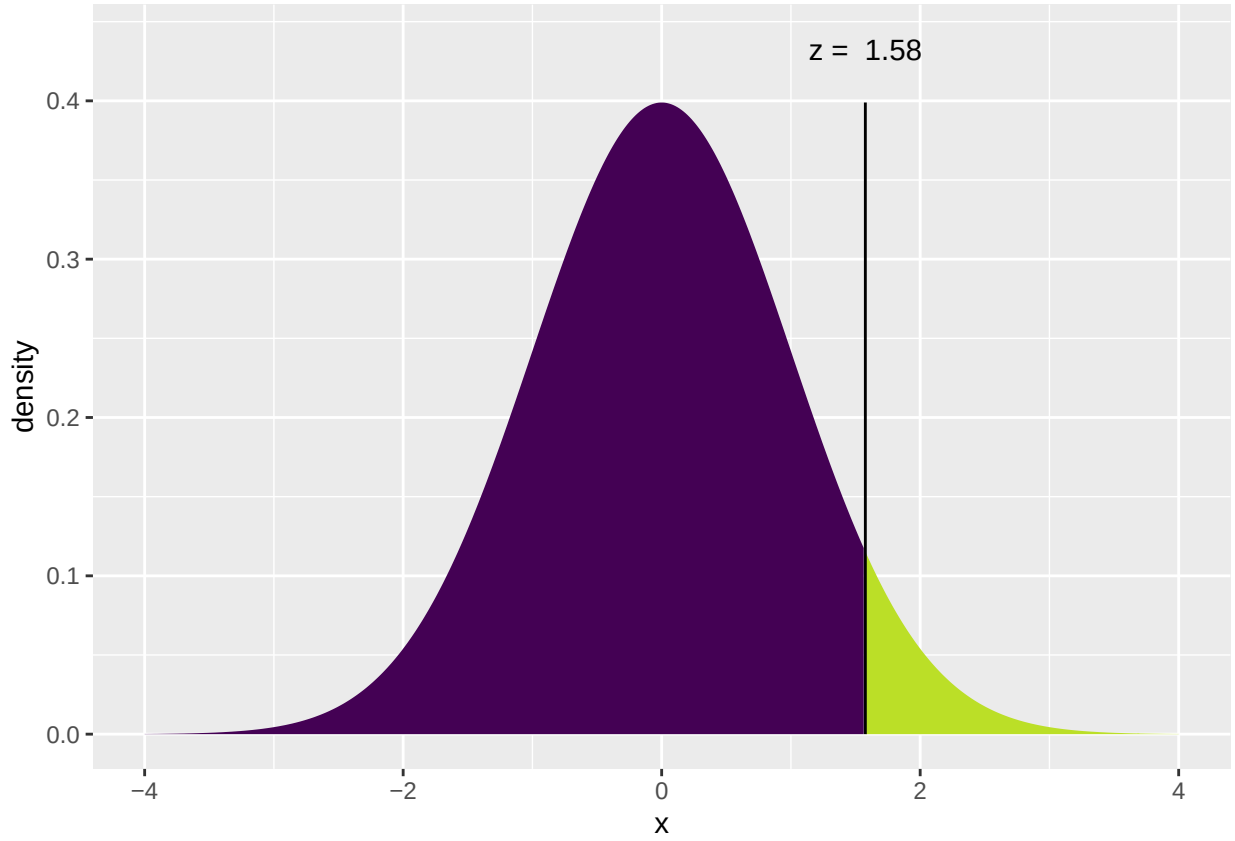
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      28
## 2 TRUE                       22
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 22/50

Canada_analysis_science %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      29
## 2 TRUE                       41
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 41/70

p_hat = (22+41)/(50+70)

sd <- sqrt((((p_hat)*(1-p_hat))/50)+(((p_hat)*(1-p_hat))/70))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.576) = P(Z <= 1.576) = 0.9425
##  P(X >  1.576) = P(Z >  1.576) = 0.05753
##
```




```
## [1] 0.1150569
```

```
#outbreak scitech
count(Canada_analysis_science, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      30
## 2 TRUE                       90
```

```
num_preoutbreak = 90
num_postoutbreak = 30
num = 120

Canada_analysis_science %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      48
## 2 TRUE                       42
```

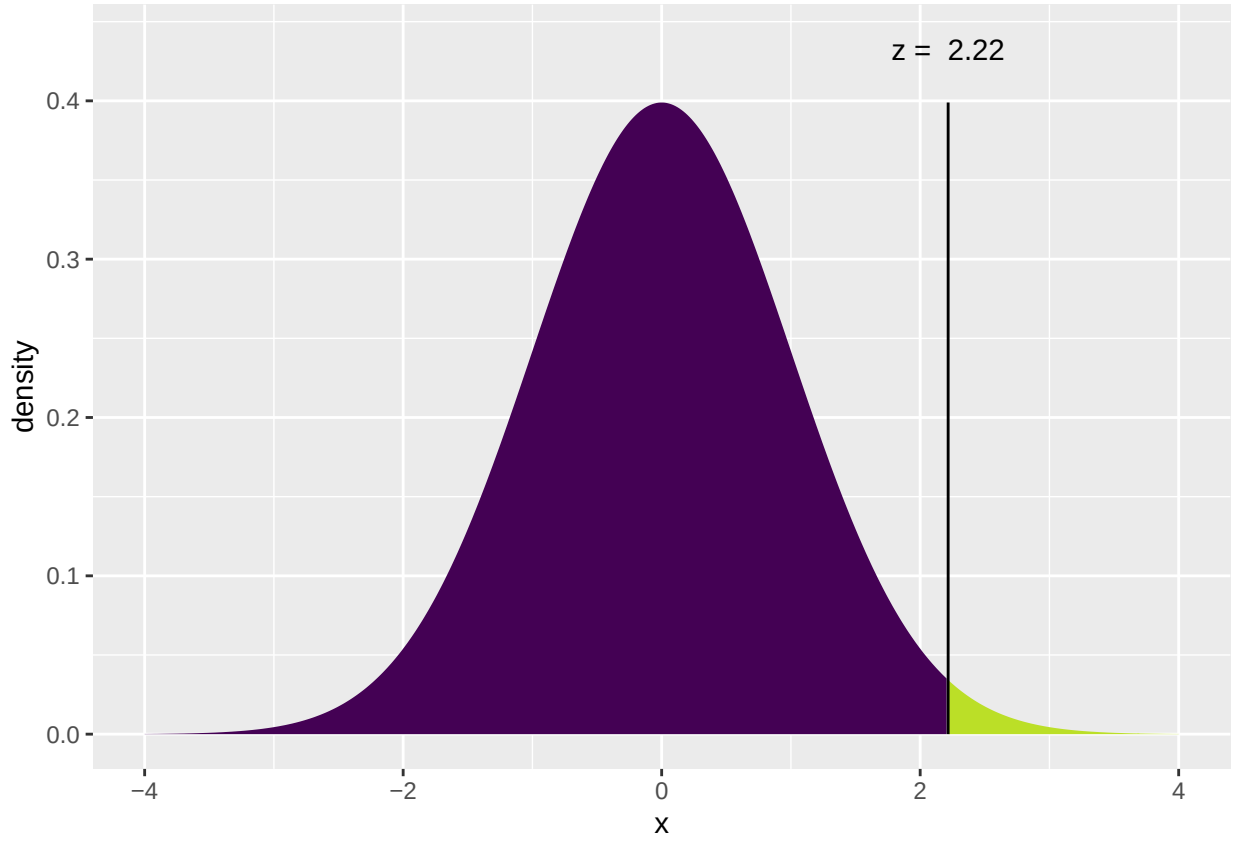
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 42/90
```

```
Canada_analysis_science %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                       9
## 2 TRUE                       21
```

```
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 21/30

p_hat = (42+21)/(90+30)

sd <- sqrt((((p_hat)*(1-p_hat))/90)+(((p_hat)*(1-p_hat))/30))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

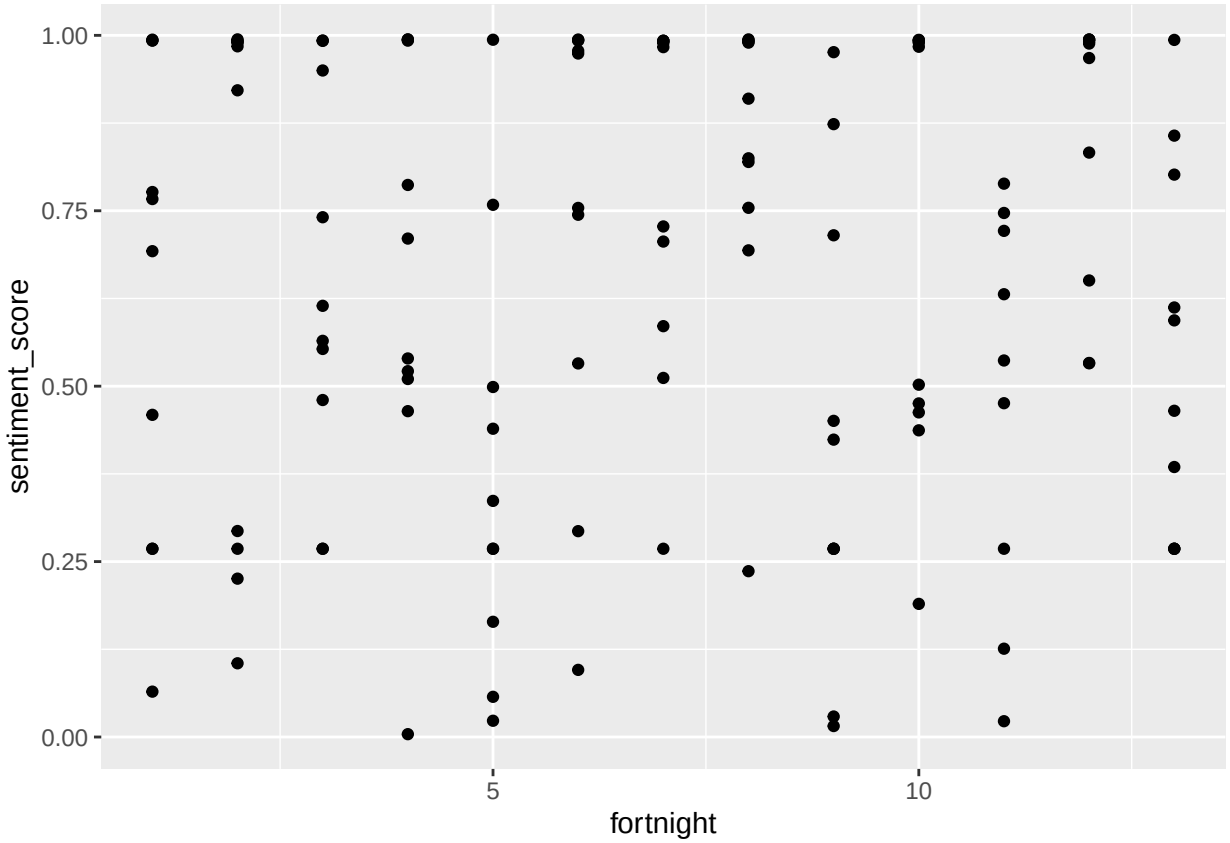
```
##
## If X ~ N(0, 1), then
##  P(X <= 2.216) = P(Z <= 2.216) = 0.9867
##  P(X >  2.216) = P(Z >  2.216) = 0.01333
##
```



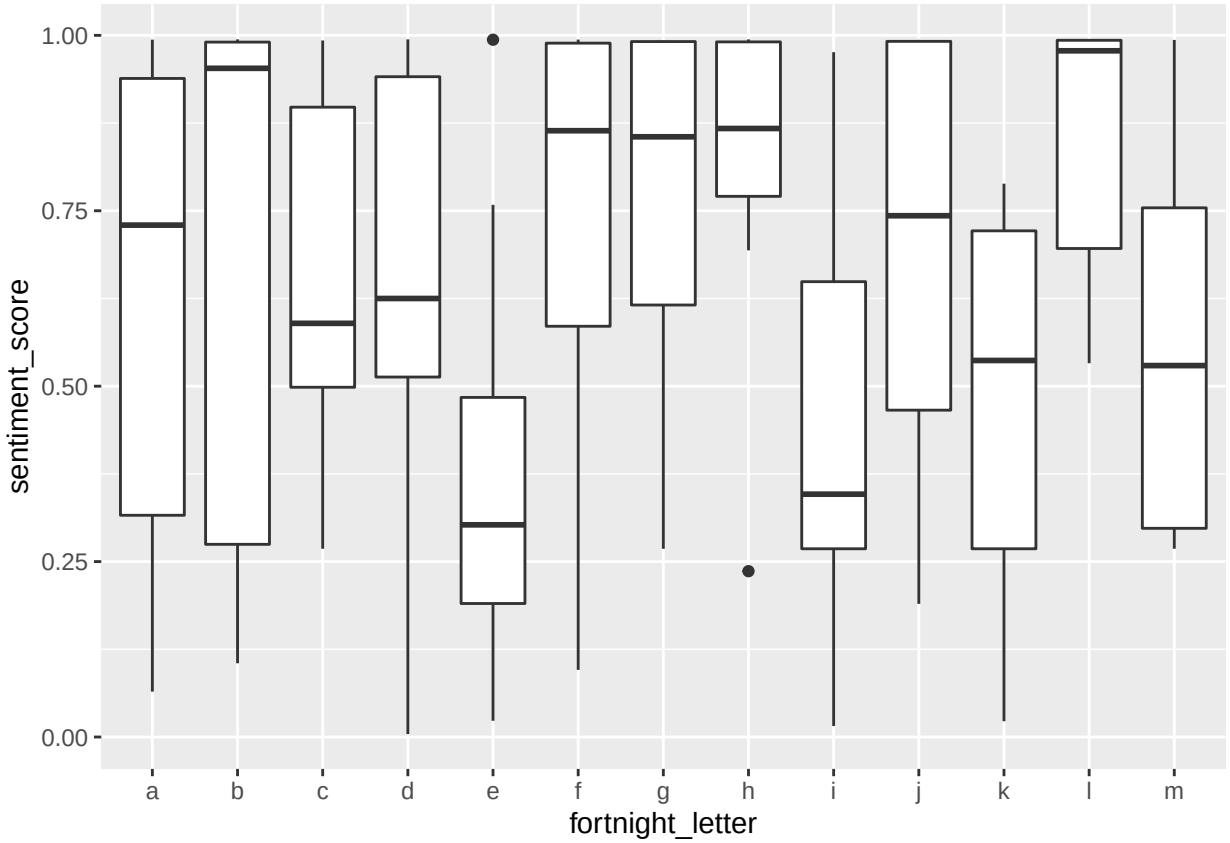

```
## [1] 0.02666641
```

```
#Youtube API All Categories
Canada_analysis_all <- Canada_analysis %>%
  filter(video_category == "All")
```

```
#data summary all categories
ggplot(Canada_analysis_all) +
  geom_point(aes(x = fortnight, y = sentiment_score))
```



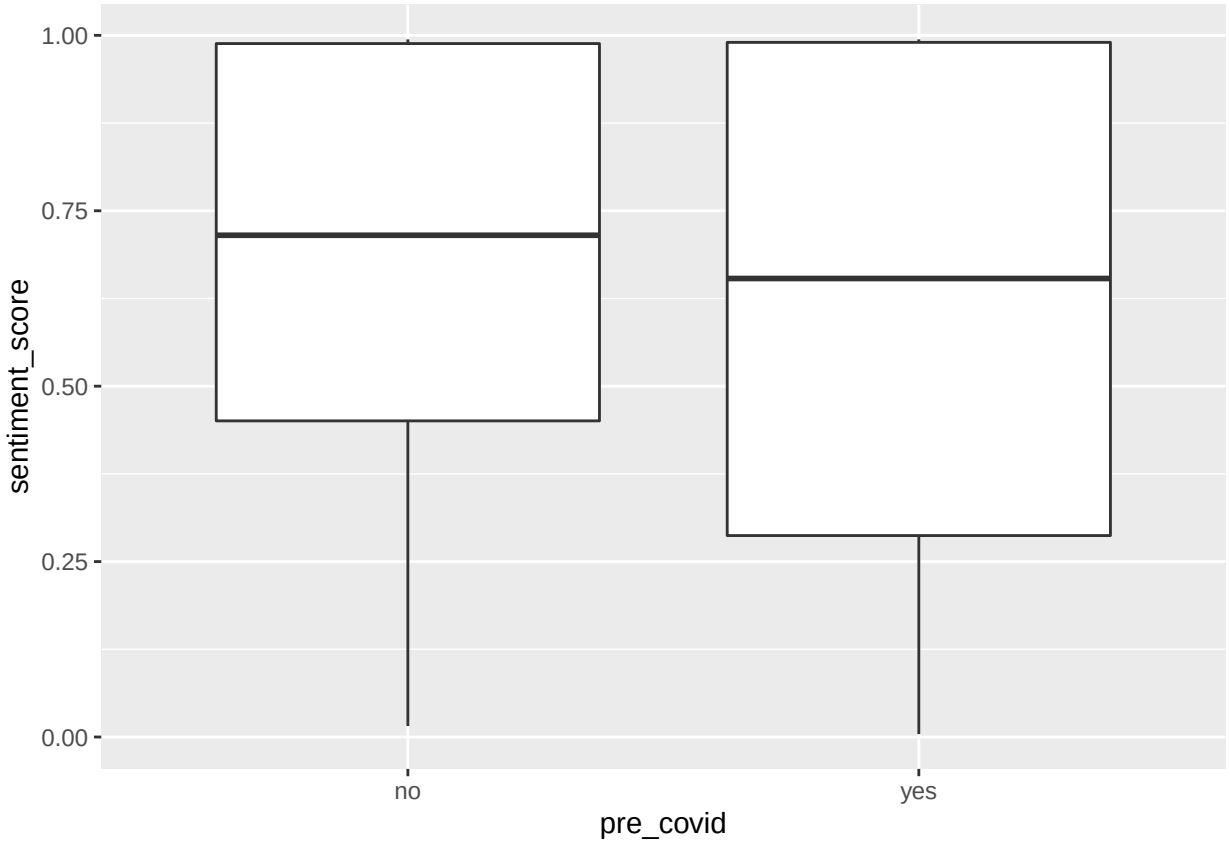

```
ggplot(Canada_analysis_all) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




```
Canada_analysis_all %>%
  group_by(fortnight) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.628
##  2         2                   0.677
##  3         3                   0.642
##  4         4                   0.652
##  5         5                   0.381
##  6         6                   0.735
##  7         7                   0.775
##  8         8                   0.821
##  9         9                   0.429
## 10        10                   0.702
## 11        11                   0.480
## 12        12                   0.848
## 13        13                   0.551
```

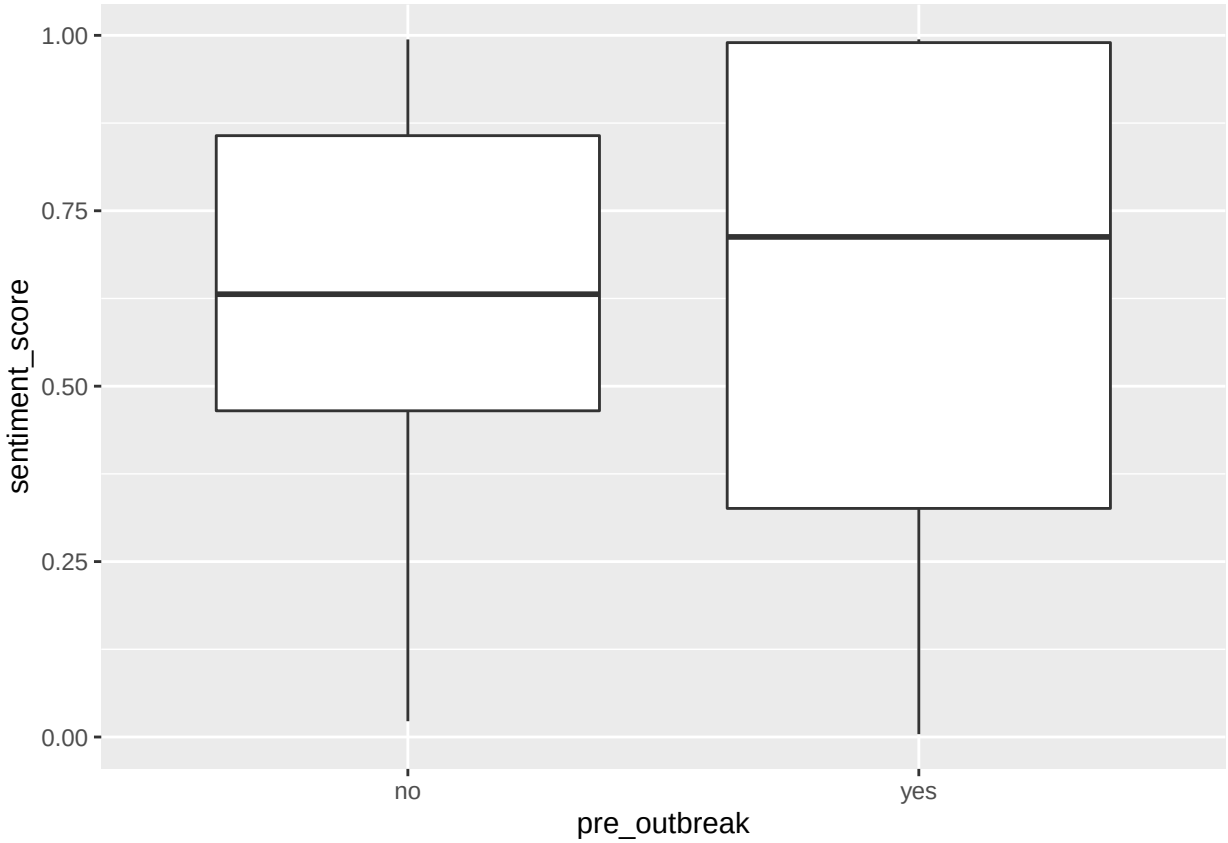
```
ggplot(Canada_analysis_all) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
Canada_analysis_all %>%
  group_by(pre_covid) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.660
## 2 yes                         0.619
```

```
ggplot(Canada_analysis_all) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```




```
Canada_analysis_all %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.631
## 2 yes                            0.644
```

```
#precovid all categories
count(Canada_analysis_all, pre_covid == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_covid == "yes"`     n
##   <lgl>                <int>
## 1 FALSE                   69
## 2 TRUE                    60
```

```
num_precovid = 60
num_postcovid = 69
num = 129

Canada_analysis_all %>%
  filter(pre_covid == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
```

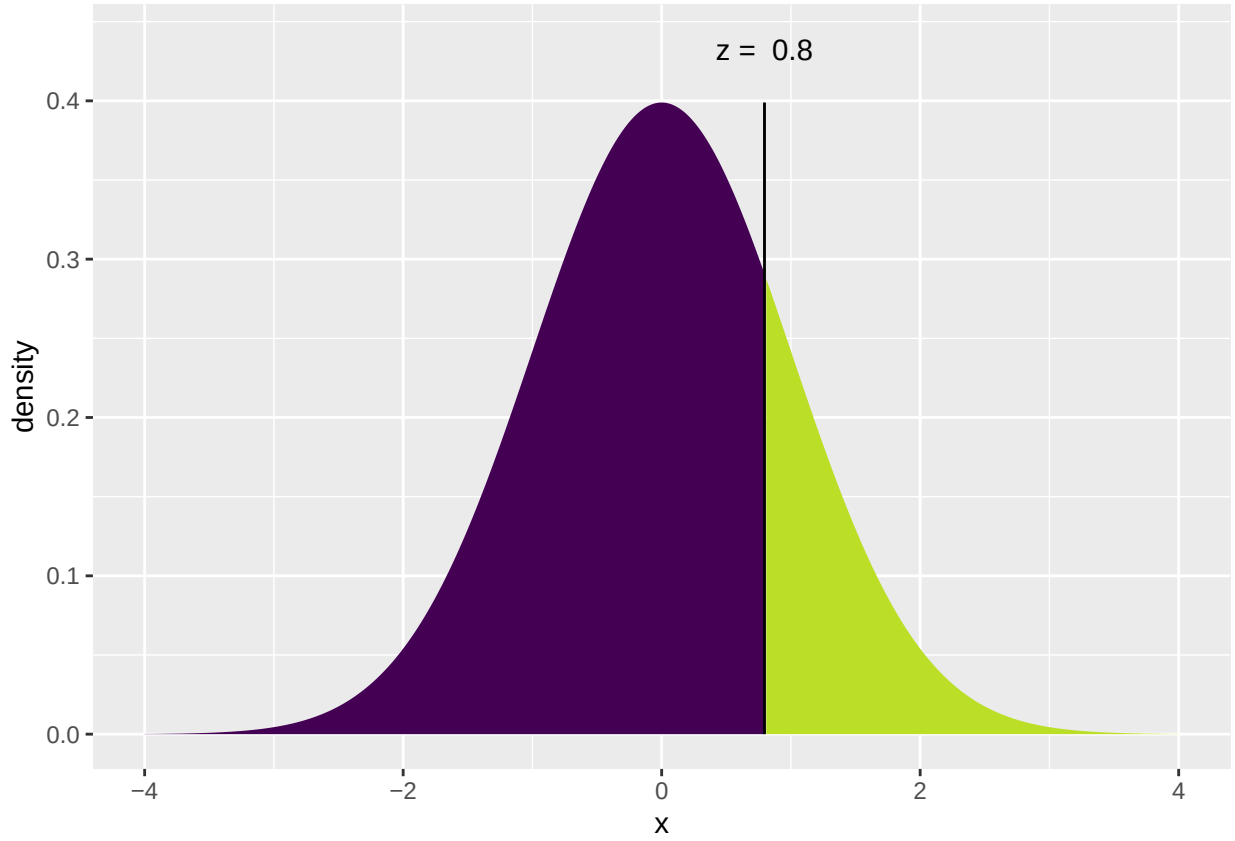
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      23
## 2 TRUE                       37
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 37/60

Canada_analysis_all %>%
  filter(pre_covid == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      22
## 2 TRUE                       47
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 47/69

p_hat = (43+47)/(60+69)

sd <- sqrt((((p_hat)*(1-p_hat))/60)+(((p_hat)*(1-p_hat))/69))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.7955) = P(Z <= 0.7955) = 0.7868
##  P(X >  0.7955) = P(Z >  0.7955) = 0.2132
##
```




```
## [1] 0.4263097
```

```
#outbreak all categories
count(Canada_analysis_all, pre_outbreak == "yes")
```

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      29
## 2 TRUE                      100
```

```
num_preoutbreak = 100
num_postoutbreak = 29
num = 129

Canada_analysis_all %>%
  filter(pre_outbreak == "yes") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      36
## 2 TRUE                       64
```

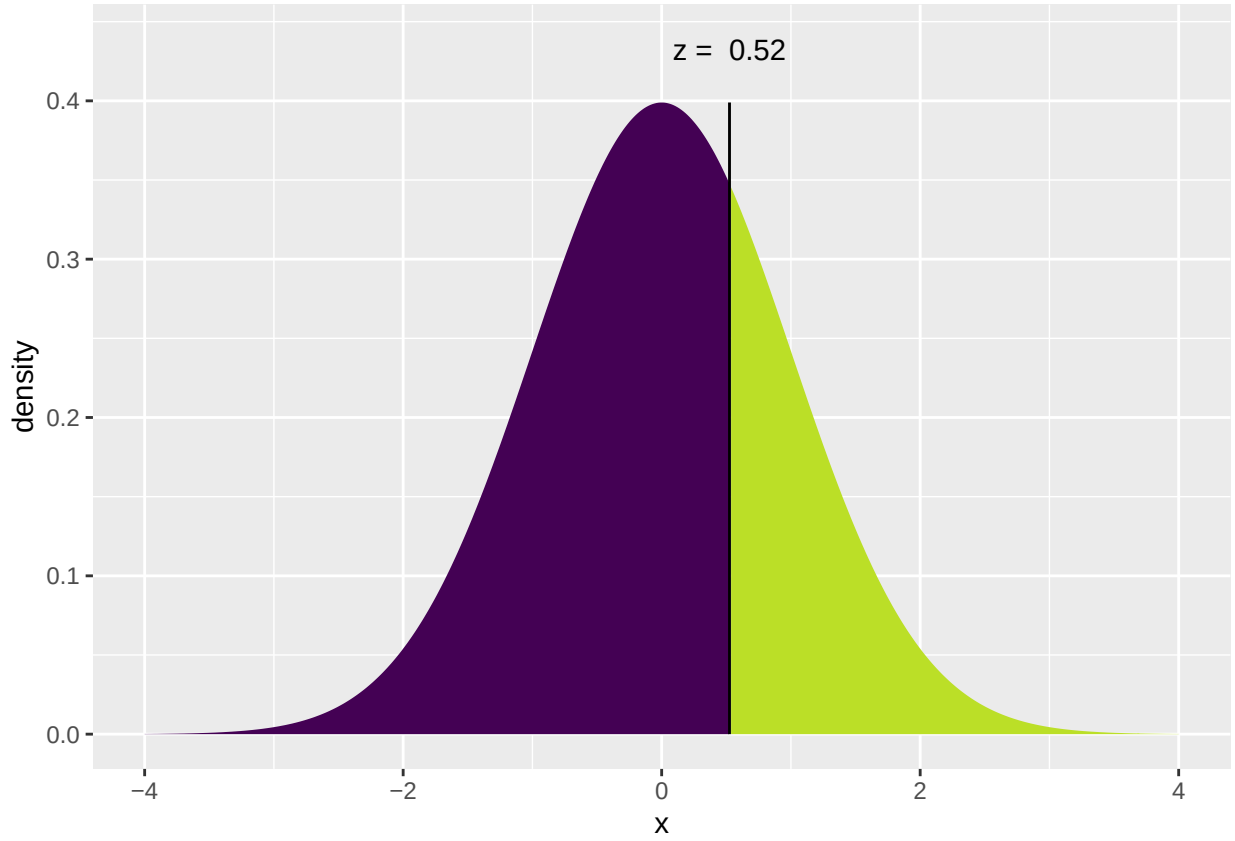
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 64/100
```

```r
Canada_analysis_all %>%
  filter(pre_outbreak == "no") %>%
  count(sentiment_score > 0.5)
```

```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                       9
## 2 TRUE                       20
```

```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 20/29

p_hat = (64+29)/(100+29)

sd <- sqrt((((p_hat)*(1-p_hat))/100)+(((p_hat)*(1-p_hat))/29))
z_score <- ((p_hat2-p_hat1)-0)/sd

#p-value
2* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.5249) = P(Z <= 0.5249) = 0.7002
##  P(X >  0.5249) = P(Z >  0.5249) = 0.2998
##
```




```
## [1] 0.5996619
```

```
#Two independent samples t-tests; Comparing two independent means

#pre_covid music
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = Canada_analysis_music)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -0.2672, df = 119.77, p-value = 0.7898
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1391216  0.1060367
## sample estimates:
##  mean in group no mean in group yes
##         0.4668994         0.4834419
```

```
#pre_outbreak music
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = Canada_analysis_music)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.33683, df = 50.658, p-value = 0.7376
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.1153392  0.1618364
## sample estimates:
##  mean in group no mean in group yes
##         0.4922807         0.4690321
```

```
#pre_covid travel and events
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = Canada_analysis_travel)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -2.2516, df = 123.62, p-value = 0.02611
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.22233944 -0.01430649
## sample estimates:
##  mean in group no mean in group yes
##         0.6093925         0.7277155
```

```
#pre_outbreak travel and events
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = Canada_analysis_travel)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -1.0651, df = 45.48, p-value = 0.2925
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.20038455  0.06173478
## sample estimates:
##  mean in group no mean in group yes
##         0.6106763         0.6800012
```

```
#pre_covid people and blogs
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = Canada_analysis_people)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.54817, df = 121.53, p-value = 0.5846
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0871798  0.1539478
## sample estimates:
##  mean in group no mean in group yes
##         0.6578484         0.6244644
```

```
#pre_outbreak people and blogs
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = Canada_analysis_people)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  sentiment_score by pre_outbreak
## t = 0.424, df = 45.208, p-value = 0.6736
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1156934  0.1774030
## sample estimates:
##  mean in group no mean in group yes
##         0.6662394         0.6353846
```

```
#pre_covid entertainment
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = Canada_analysis_entertainment)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -0.41492, df = 125.02, p-value = 0.6789
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.14577412  0.09524471
## sample estimates:
##  mean in group no mean in group yes
##         0.6189941         0.6442588
```

```
#pre_outbreak entertainment
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = Canada_analysis_entertainment)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.070074, df = 43.431, p-value = 0.9445
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1507601  0.1616174
## sample estimates:
##  mean in group no mean in group yes
##         0.6347959         0.6293672
```

```
#pre_covid news and politics
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = Canada_analysis_news)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -1.0656, df = 118.88, p-value = 0.2888
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.17468739  0.05245006
## sample estimates:
##  mean in group no mean in group yes
##         0.5375949         0.5987135
```

```
#pre_outbreak news and politics
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = Canada_analysis_news)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -1.6291, df = 57.176, p-value = 0.1088
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.21914523  0.02252167
## sample estimates:
##  mean in group no mean in group yes
##         0.4900998         0.5884116
```

```
#pre_covid how-to and style
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = Canada_analysis_how_to)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -2.123, df = 120.65, p-value = 0.0358
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.238966517 -0.008340396
## sample estimates:
##  mean in group no mean in group yes
##         0.5624608         0.6861143
```

```
#pre_outbreak how-to and style
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = Canada_analysis_how_to)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -3.0012, df = 52.365, p-value = 0.004113
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.31987390 -0.06355277
## sample estimates:
##  mean in group no mean in group yes
##         0.4728452         0.6645585
```

```
#pre_covid education
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = Canada_analysis_education)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.3226, df = 116.86, p-value = 0.1885
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03776761  0.18962897
## sample estimates:
##  mean in group no mean in group yes
##         0.6382658         0.5623351
```

```r
#pre_outbreak education
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = Canada_analysis_education)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 1.273, df = 47.259, p-value = 0.2093
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04993426  0.22208546
## sample estimates:
##  mean in group no mean in group yes
##         0.6694329         0.5833573
```

```r
#pre_covid science and technology
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = Canada_analysis_science)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.3656, df = 103.54, p-value = 0.175
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03980387  0.21585451
## sample estimates:
##  mean in group no mean in group yes
##         0.6009992         0.5129738
```

```r
#pre_outbreak science and technology
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = Canada_analysis_science)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 1.9712, df = 49.931, p-value = 0.05426
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.002706989  0.287478040
## sample estimates:
##  mean in group no mean in group yes
##         0.6711111         0.5287256
```

```r
#pre_covid all categories
t.test(sentiment_score ~ pre_covid, alternative = "two.sided", data = Canada_analysis_all)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.73244, df = 120.59, p-value = 0.4653
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07046957  0.15322603
```

```
## sample estimates:
##  mean in group no mean in group yes
##         0.6604291         0.6190509
```

```
#pre_outbreak categories
t.test(sentiment_score ~ pre_outbreak, alternative = "two.sided", data = Canada_analysis_all)
```

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.20242, df = 50.364, p-value = 0.8404
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1392708  0.1137657
## sample estimates:
##  mean in group no mean in group yes
##         0.6312978         0.6440503
```