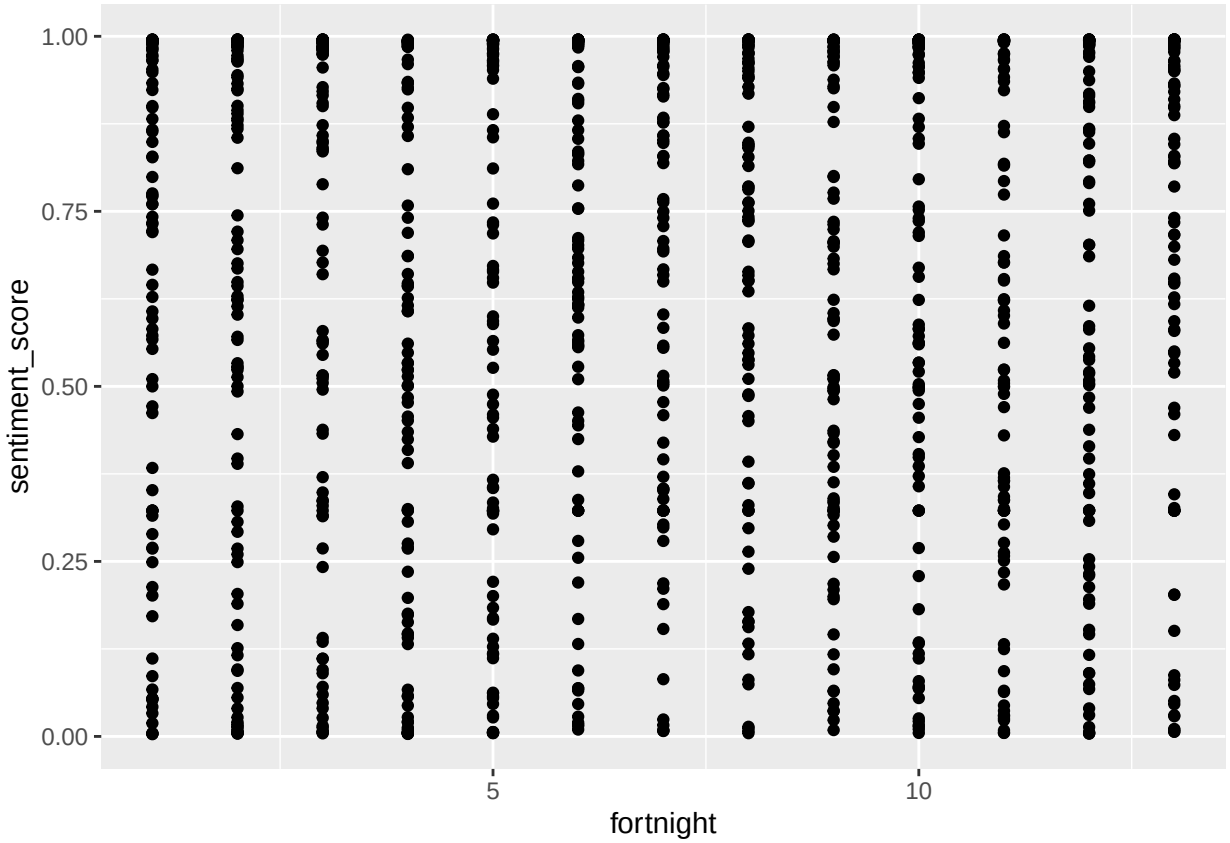
Datafest Data Analysis

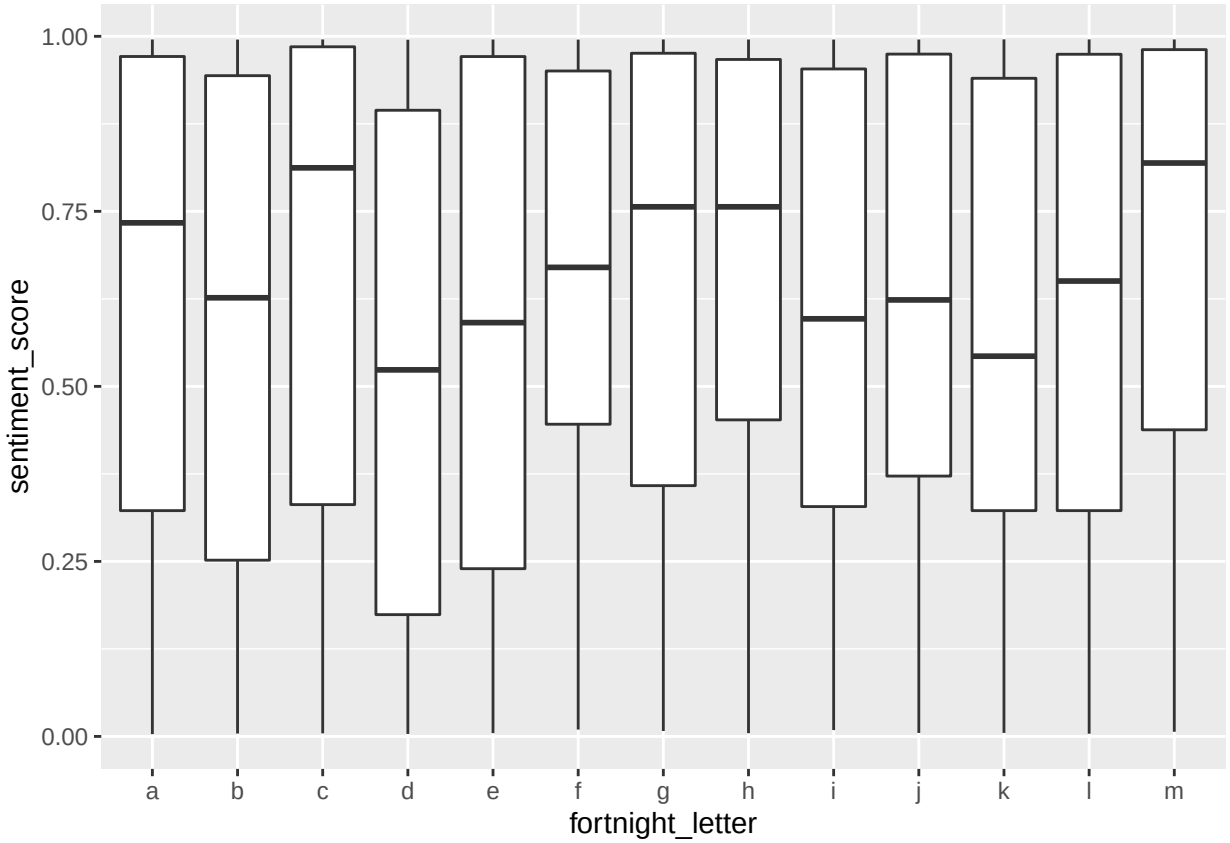
```
#US analysis import
US_analysis = read_excel("C:\\Users\\gtham\\OneDrive - Pomona College\\A - DATAFEST\\Analysis Datasets\\
```

```
US_analysis_music <- US_analysis %>%
  filter(video_category == "Music")

US_analysis_travel <- US_analysis %>%
  filter(video_category == "Travel and Events")

US_analysis_people <- US_analysis %>%
  filter(video_category == "People and Blogs")

US_analysis_entertainment <- US_analysis %>%
  filter(video_category == "Entertainment")

US_analysis_news <- US_analysis %>%
  filter(video_category == "News and Politics")

US_analysis_how_to <- US_analysis %>%
  filter(video_category == "How-to and Style")

US_analysis_education <- US_analysis %>%
  filter(video_category == "Education")

US_analysis_science <- US_analysis %>%
  filter(video_category == "Science and Technology")
```

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



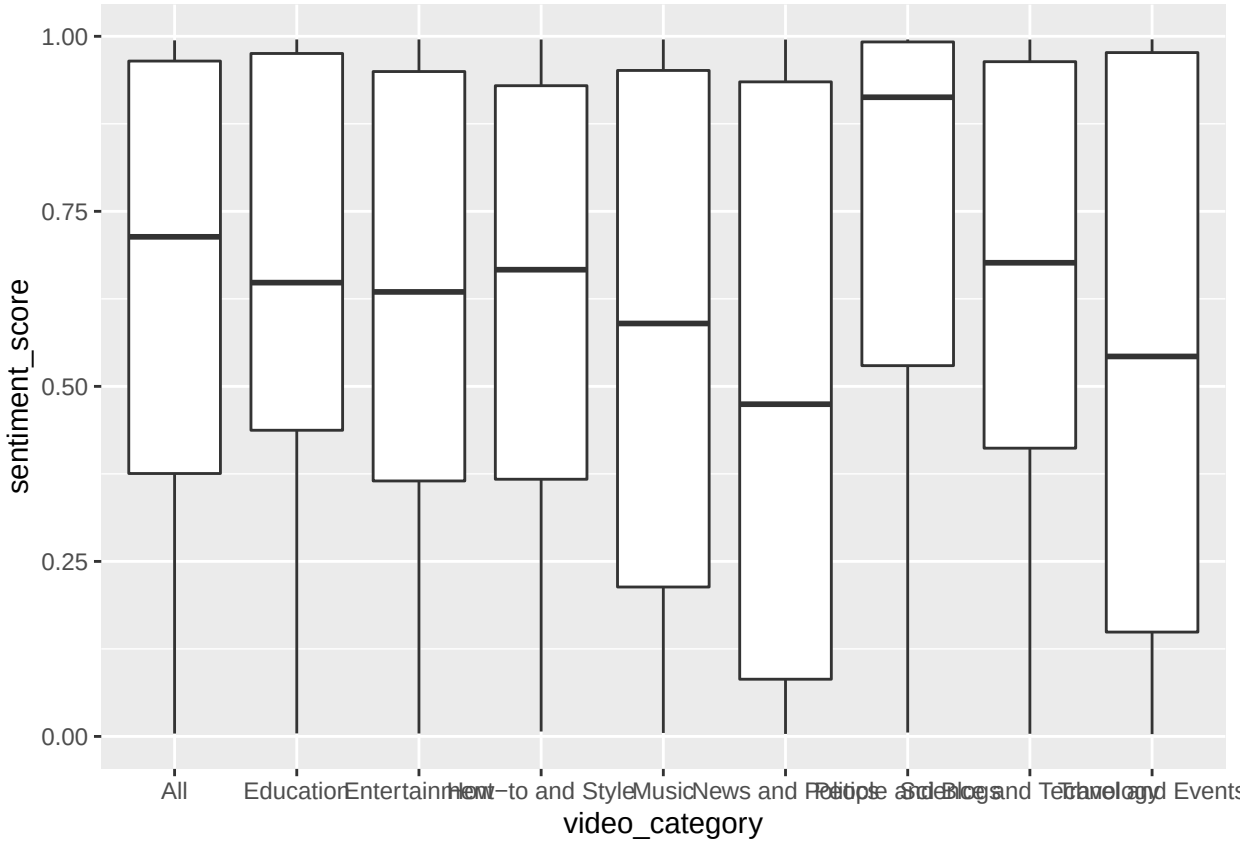

```
ggplot(US_analysis) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.625
##  2         2                   0.579
##  3         3                   0.642
##  4         4                   0.518
##  5         5                   0.569
##  6         6                   0.642
##  7         7                   0.667
##  8         8                   0.669
##  9         9                   0.596
## 10        10                   0.617
## 11        11                   0.562
## 12        12                   0.606
## 13        13                   0.675
```

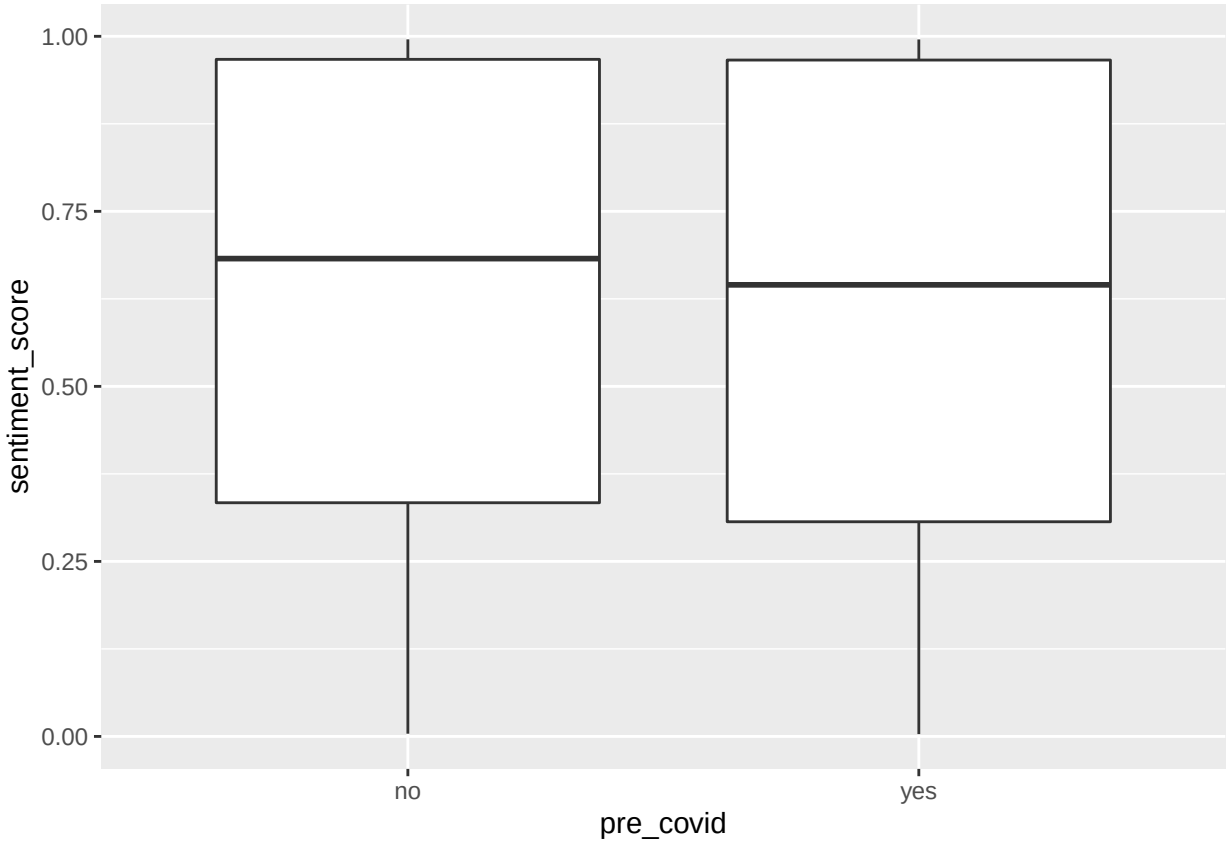
```
ggplot(US_analysis) +
  geom_boxplot(aes(x = video_category, y = sentiment_score))
```




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

```
## # A tibble: 9 x 2
##   video_category         `mean(sentiment_score)`
##   <chr>                                    <dbl>
## 1 All                                      0.645
## 2 Education                                0.649
## 3 Entertainment                            0.612
## 4 How-to and Style                         0.623
## 5 Music                                    0.564
## 6 News and Politics                        0.503
## 7 People and Blogs                         0.750
## 8 Science and Technology                   0.635
## 9 Travel and Events                        0.534
```

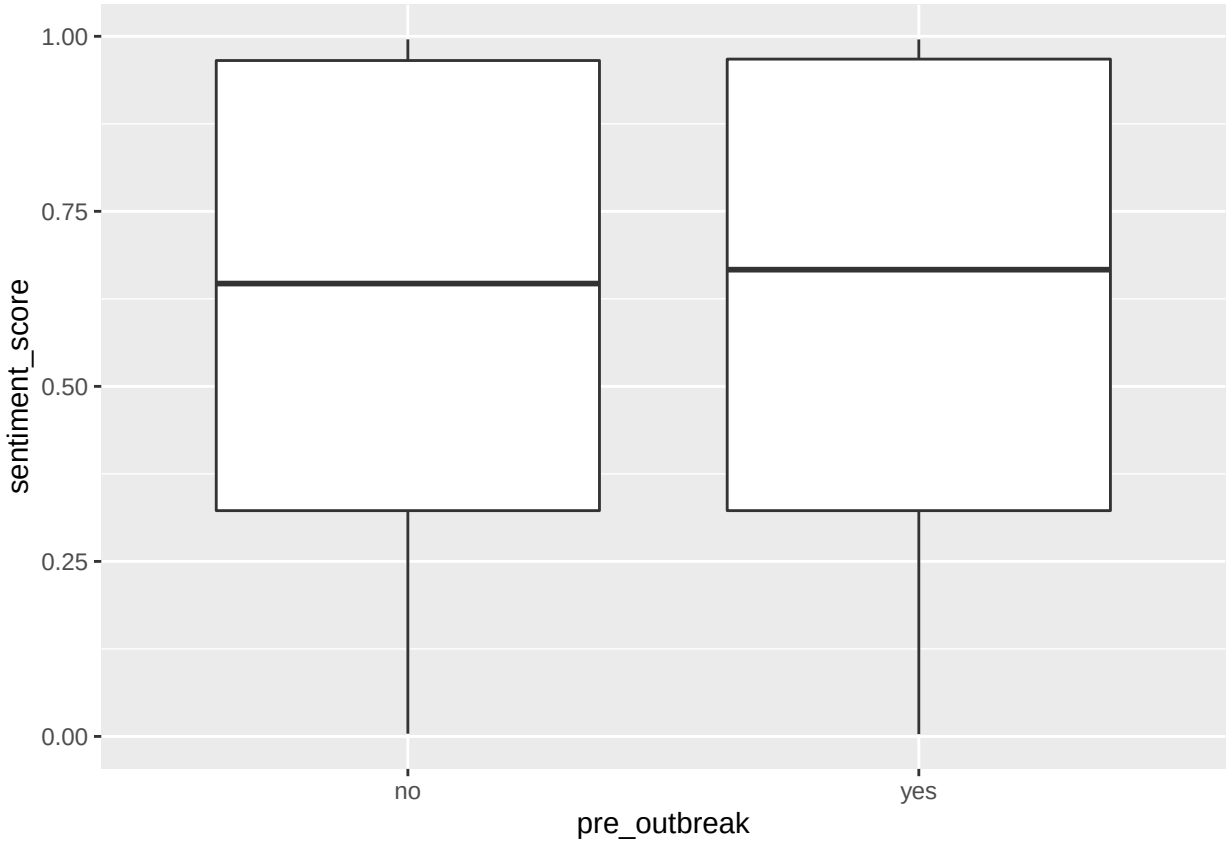
```
ggplot(US_analysis) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.628
## 2 yes                         0.596
```

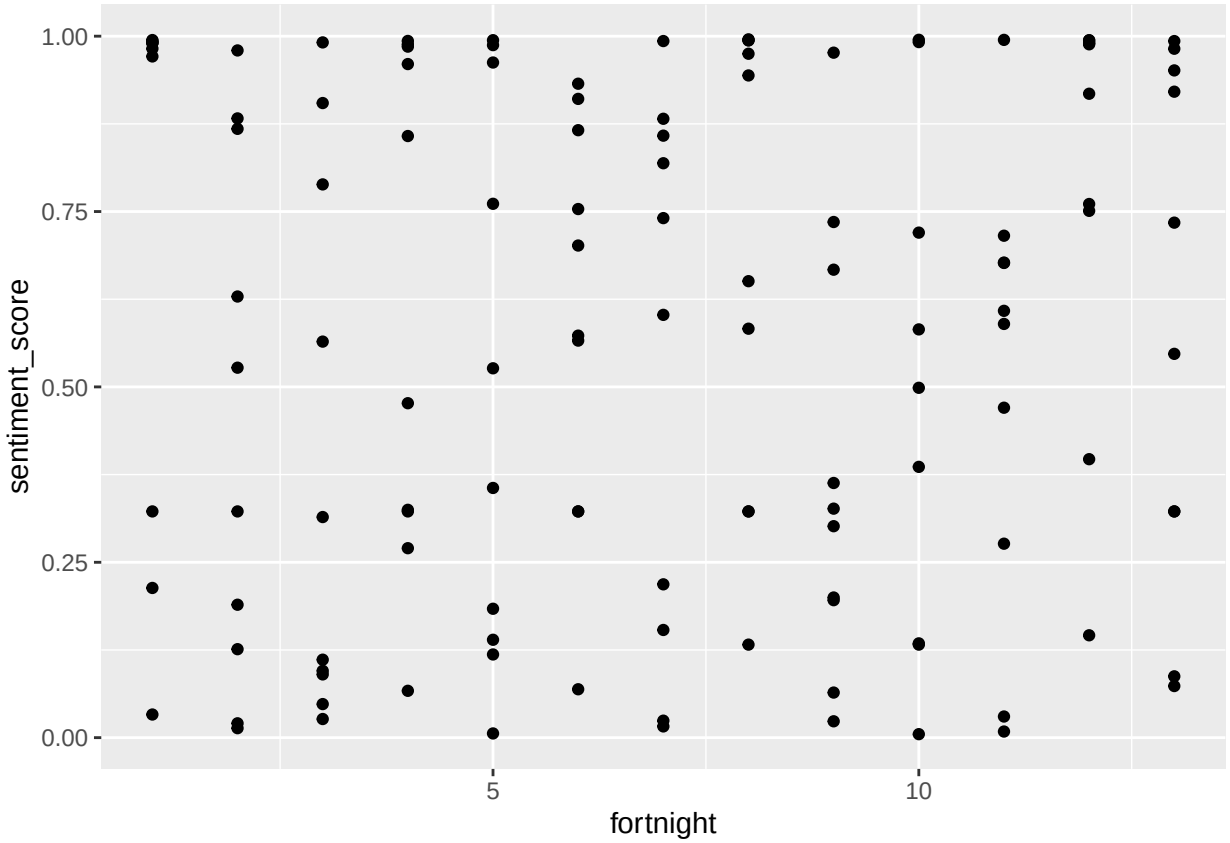
```
ggplot(US_analysis) +
  geom_boxplot(aes(x = pre_outbreak, y = sentiment_score))
```



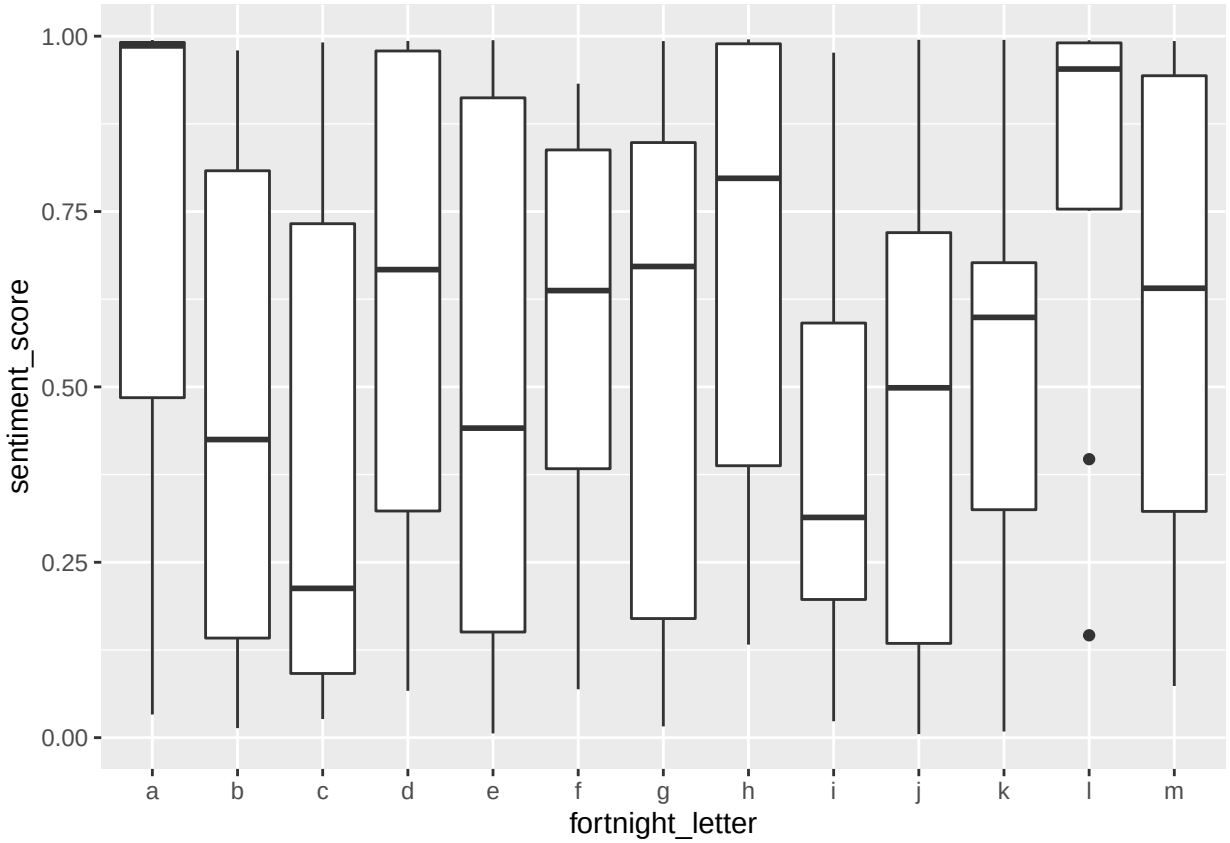

```
US_analysis %>%
  group_by(pre_outbreak) %>%
  summarize(mean(sentiment_score))
```

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.611
## 2 yes                            0.614
```

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



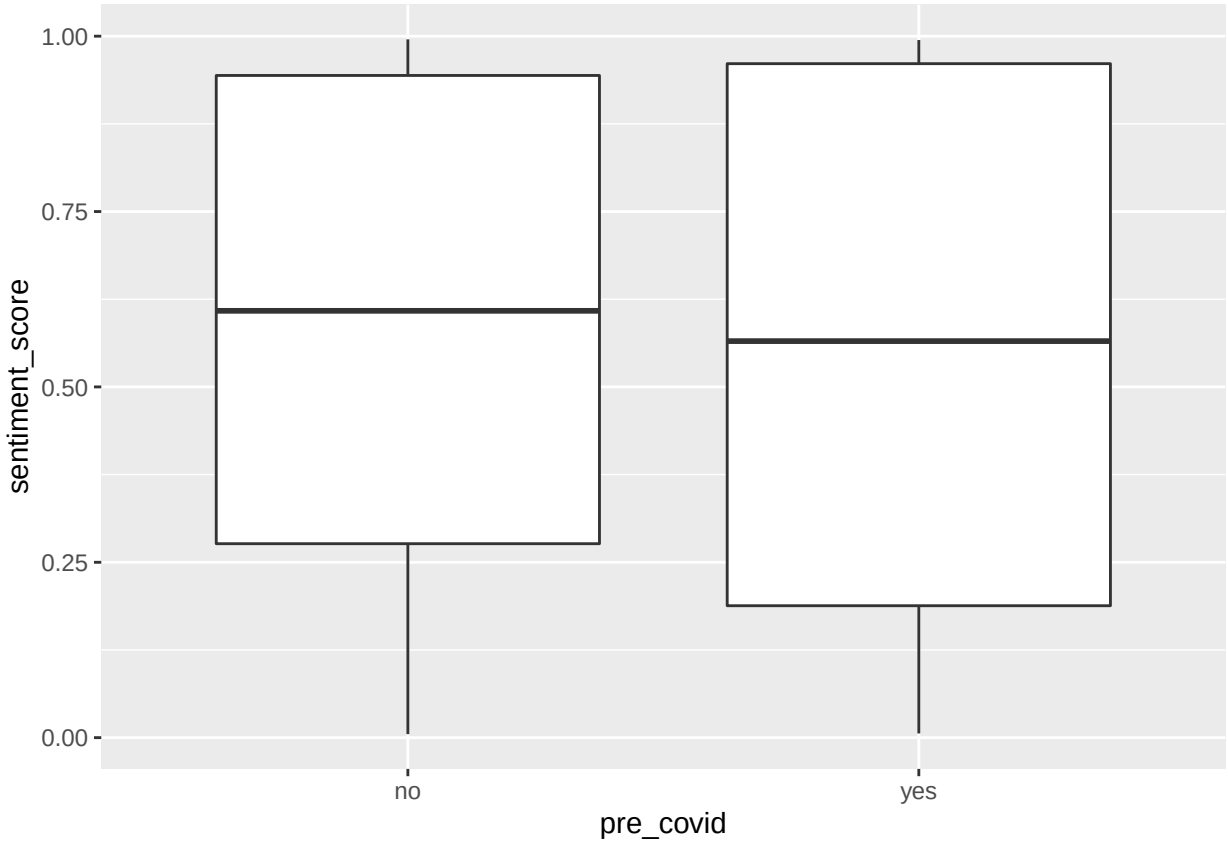

```
ggplot(US_analysis_music) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.748
##  2         2                   0.456
##  3         3                   0.393
##  4         4                   0.625
##  5         5                   0.504
##  6         6                   0.602
##  7         7                   0.531
##  8         8                   0.691
##  9         9                   0.385
## 10        10                   0.494
## 11        11                   0.505
## 12        12                   0.793
## 13        13                   0.593
```

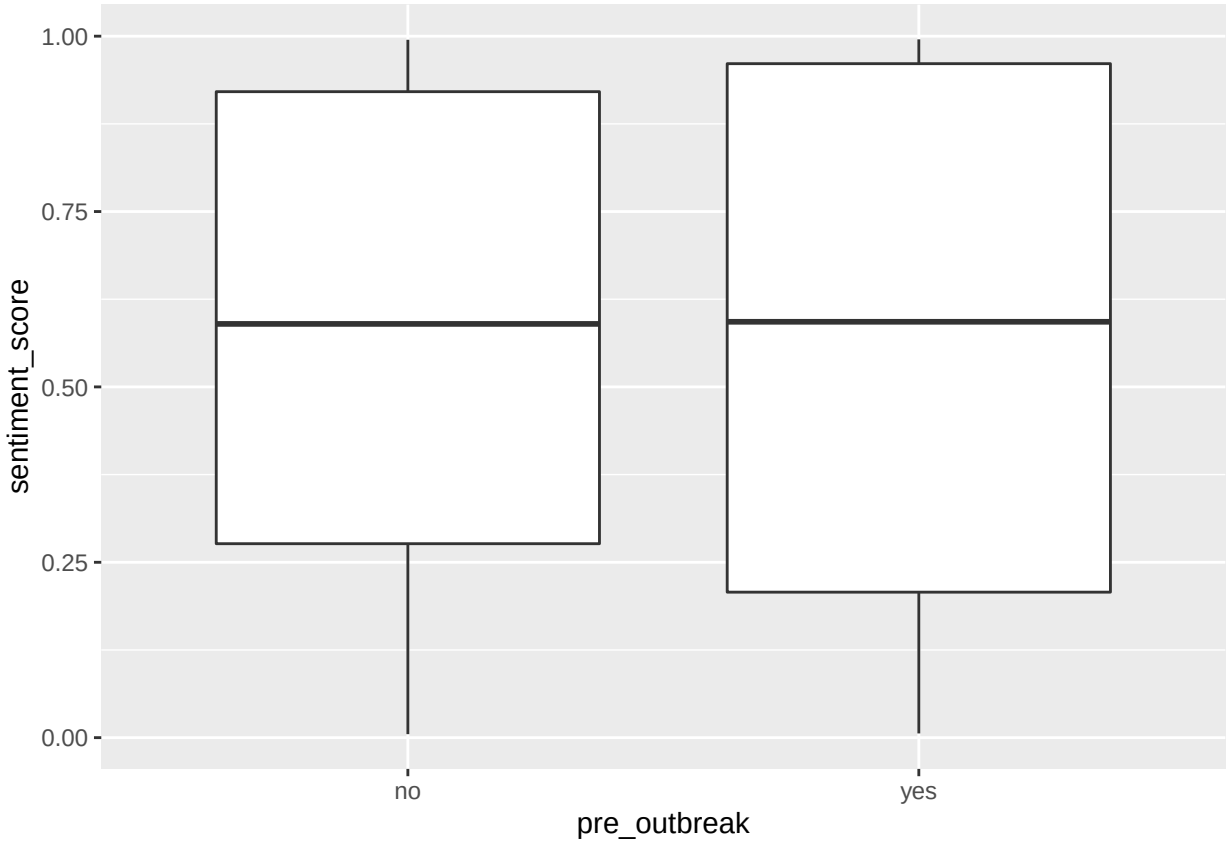
```
ggplot(US_analysis_music) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.571
## 2 yes                         0.555
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.555
## 2 yes                            0.569
```

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

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

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

```
m_num_precovid = 60
m_num_postcovid = 69
m_num = 129

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

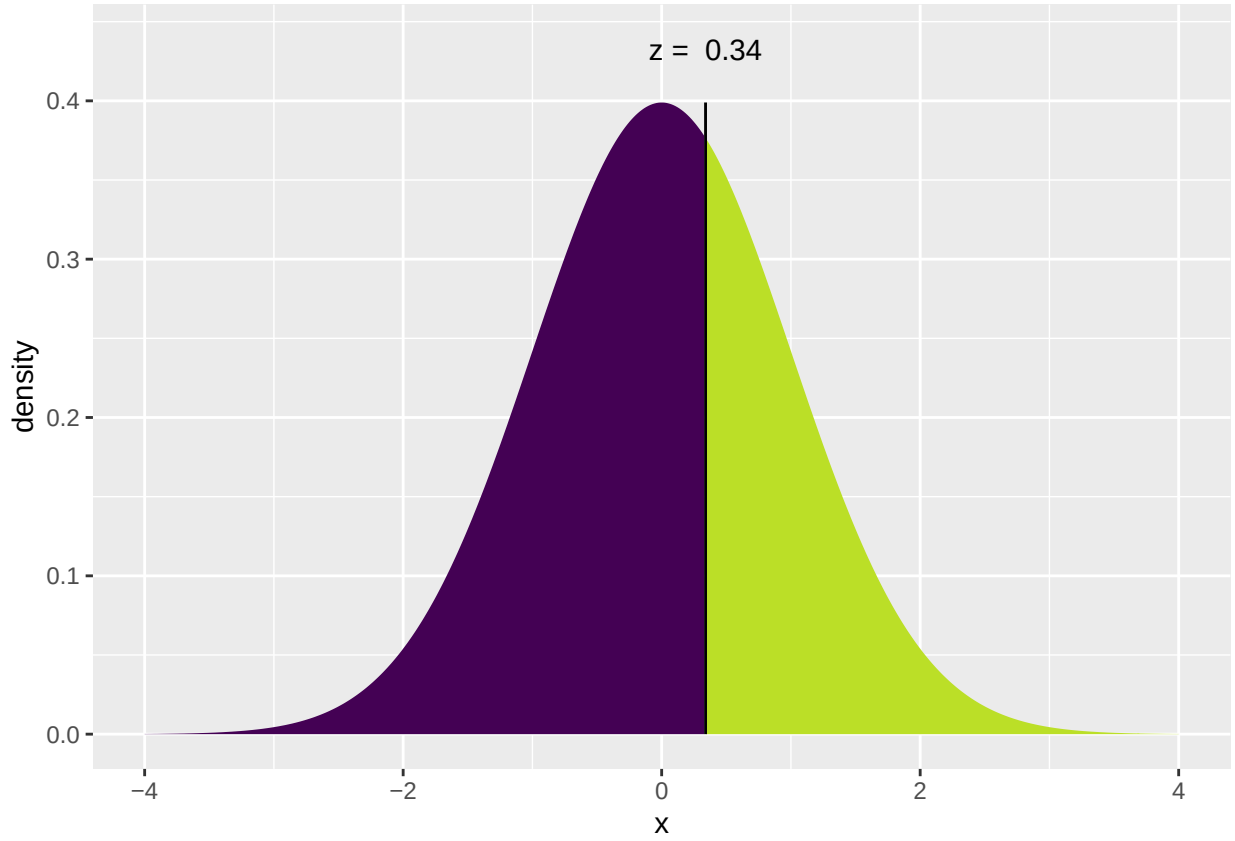
```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      27
## 2 TRUE                       33
```

```
p_hat_1_m_pos = 33/60

US_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                      29
## 2 TRUE                       40
```

```
p_hat_2_m_pos = 40/69

p_hat_m_pos = (33+40)/(60+69)

sd <- sqrt((((p_hat_m_pos)*(1-p_hat_m_pos))/60)+(((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* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.3396) = P(Z <= 0.3396) = 0.6329
##  P(X >  0.3396) = P(Z >  0.3396) = 0.3671
##
```




```
## [1] 0.7341715
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      49
## 2 TRUE                       80
```

```
m_num_preoutbreak = 80
m_num_postoutbreak = 49
m_num = 129

US_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                      34
## 2 TRUE                       46
```

```
p_hat_1_m_pos = 46/80

US_analysis_music %>%
```

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

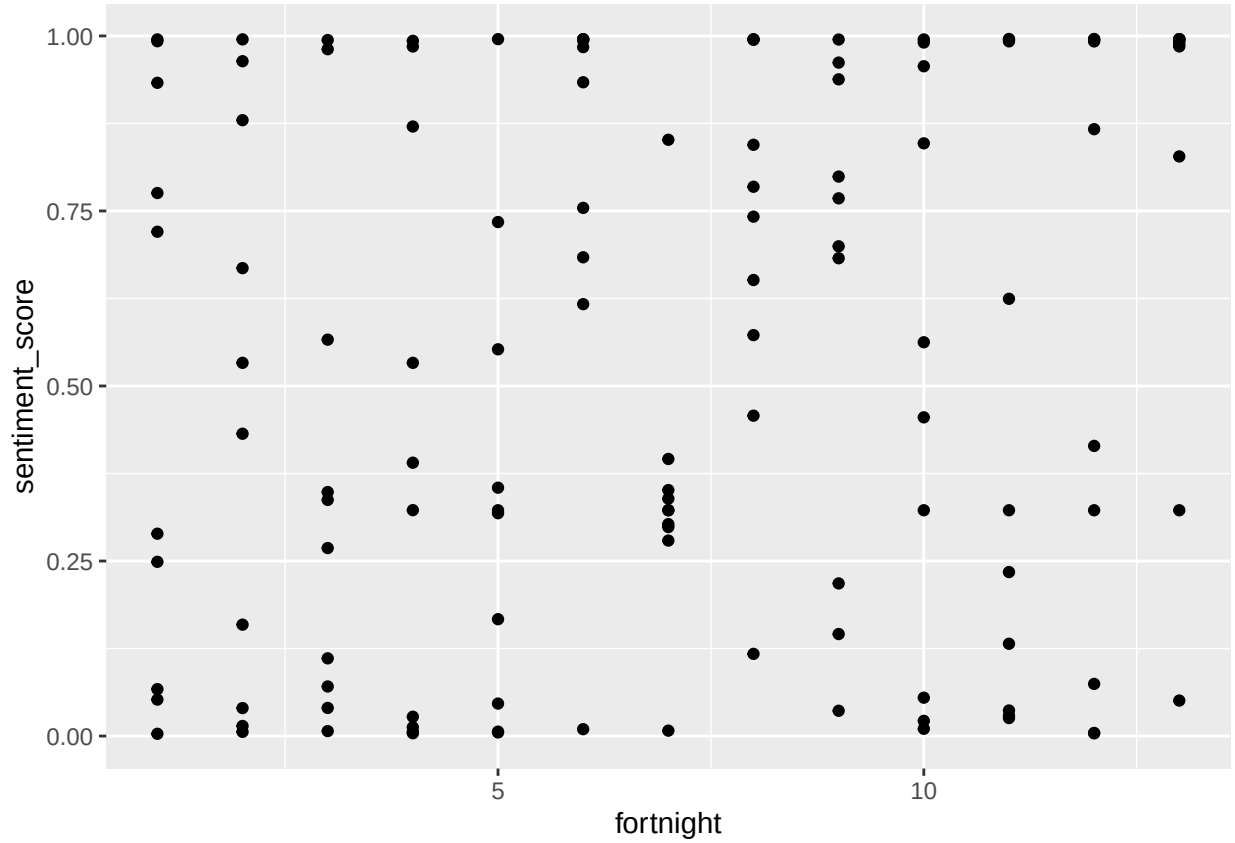
```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      22
## 2 TRUE                       27
```

```
p_hat_2_m_pos = 27/49

p_hat_m_pos = (46+27)/(80+49)

sd <- sqrt((((p_hat_m_pos)*(1-p_hat_m_pos))/80)+(((p_hat_m_pos)*(1-p_hat_m_pos))/49))
z_score <- ((p_hat_2_m_pos-p_hat_1_m_pos)-0)/sd

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

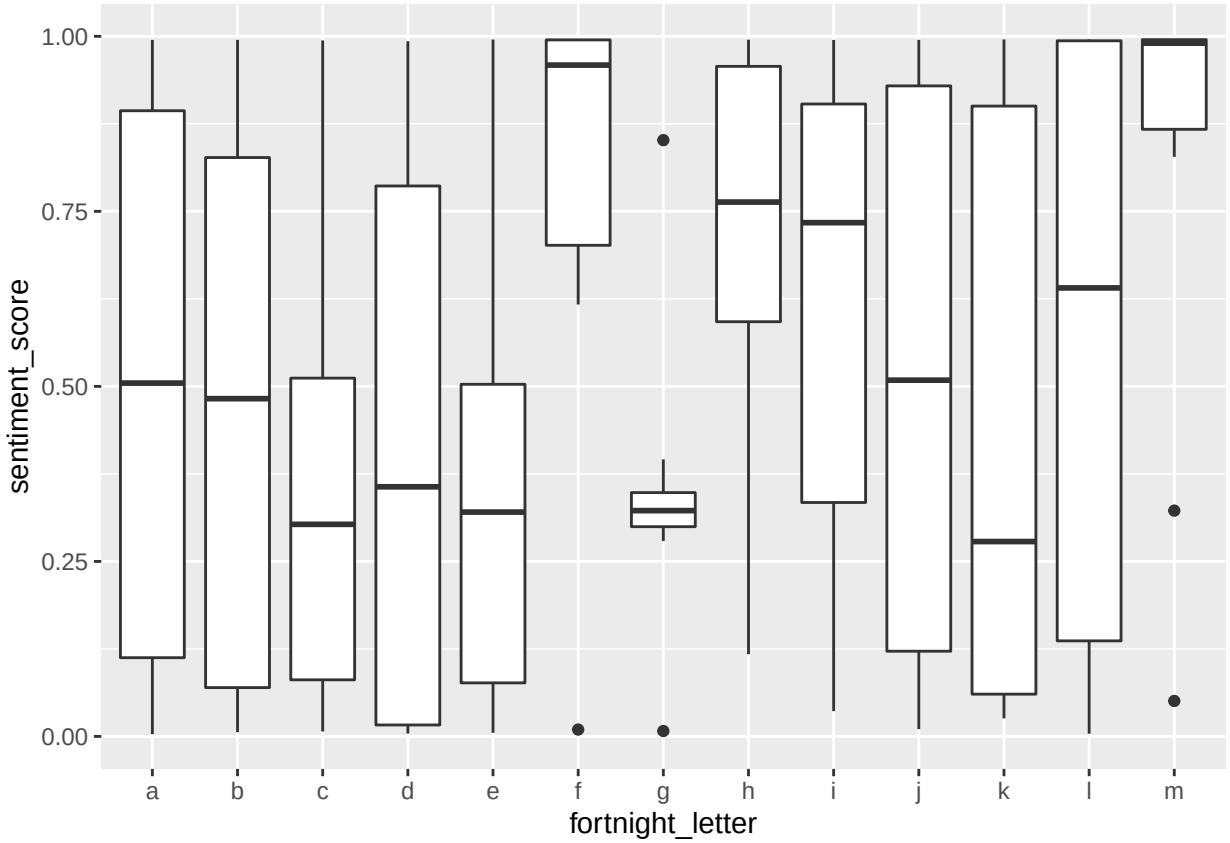
```
##
## If X ~ N(0, 1), then
##  P(X <= -0.2667) = P(Z <= -0.2667) = 0.3948
##  P(X >  -0.2667) = P(Z >  -0.2667) = 0.6052
##
```

```
## [1] 0.7896996
```

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



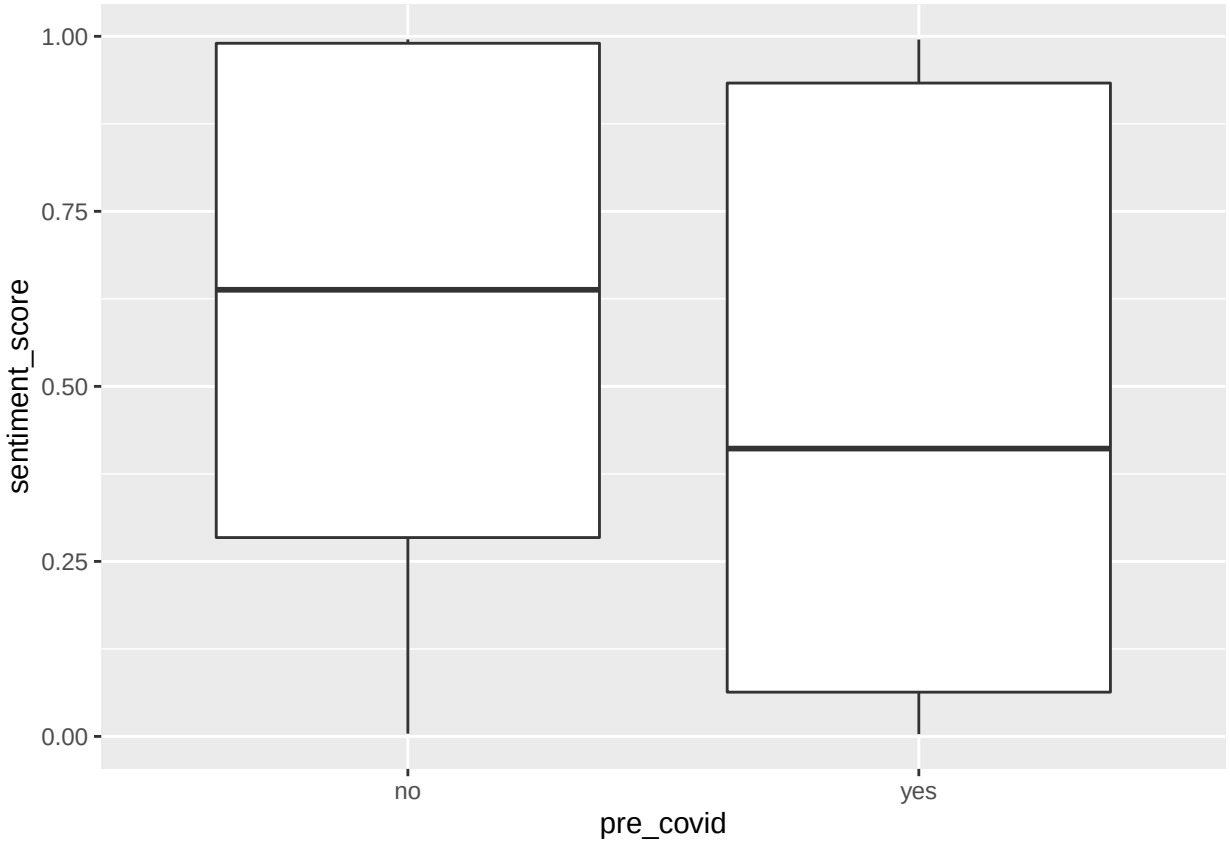

```
ggplot(US_analysis_travel) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.508
##  2         2                   0.469
##  3         3                   0.372
##  4         4                   0.414
##  5         5                   0.350
##  6         6                   0.796
##  7         7                   0.347
##  8         8                   0.715
##  9         9                   0.624
## 10        10                   0.522
## 11        11                   0.439
## 12        12                   0.566
## 13        13                   0.815
```

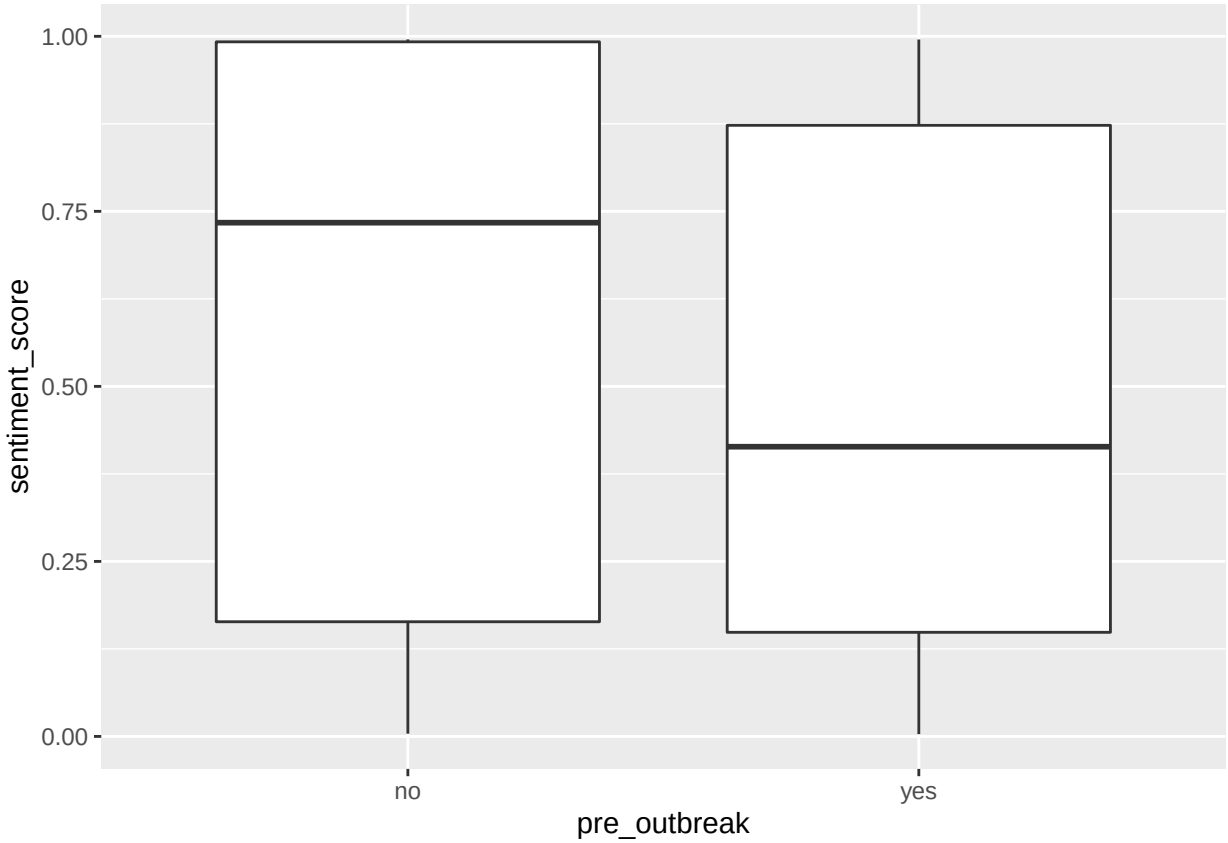
```
ggplot(US_analysis_travel) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.575
## 2 yes                         0.485
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.593
## 2 yes                            0.497
```

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

count(US_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

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

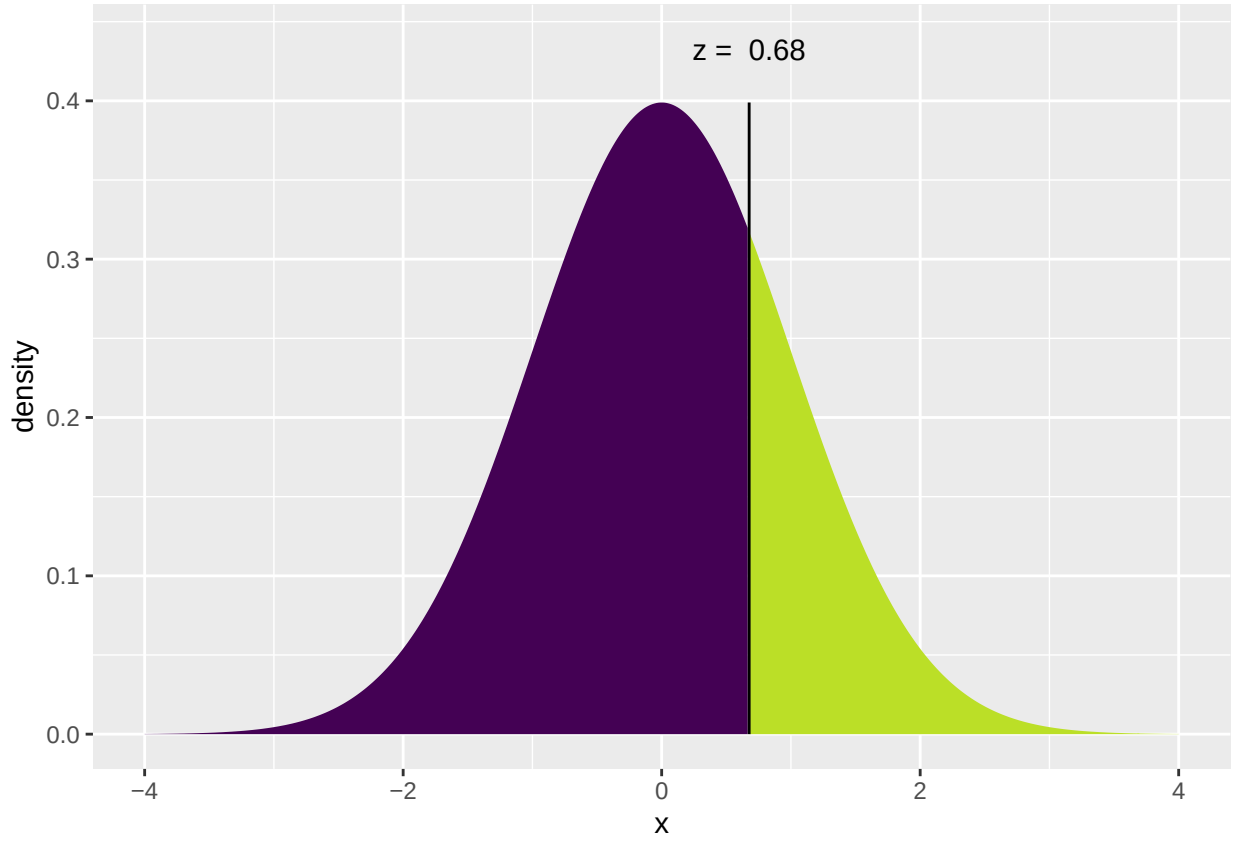
```
## # A tibble: 2 x 2
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      31
## 2 TRUE                       29
```

```
p_hat_1_t_pos = 29/60

US_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                      32
## 2 TRUE                       38
```

```
p_hat_2_t_pos = 38/70

p_hat_t_pos = (29+38)/(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* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.677) = P(Z <= 0.677) = 0.7508
##  P(X >  0.677) = P(Z >  0.677) = 0.2492
##
```




```
## [1] 0.4984152
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      50
## 2 TRUE                       80
```

```
t_num_preoutbreak = 80
t_num_postoutbreak = 50
t_num = 130

US_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                      42
## 2 TRUE                       38
```

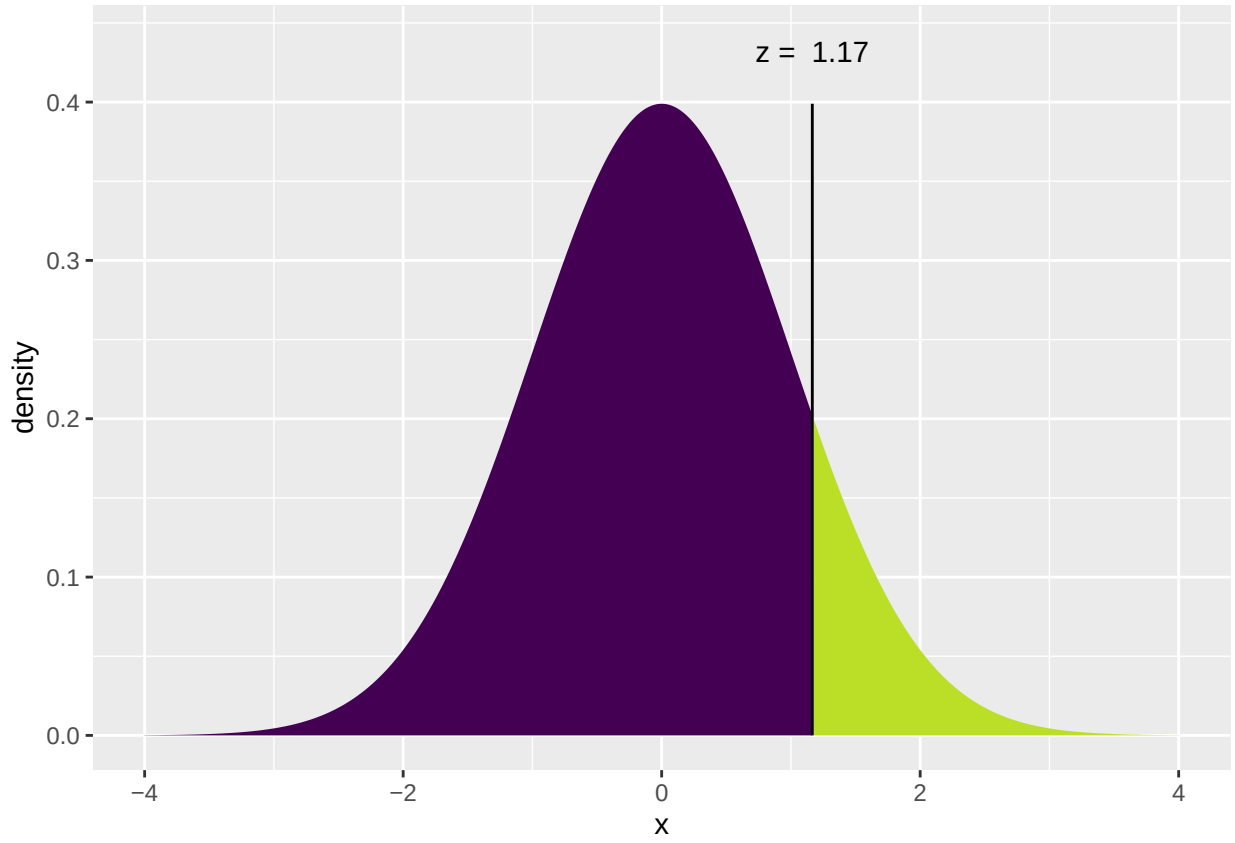
```
p_hat_1_t_pos = 38/80

US_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                      21
## 2 TRUE                       29
```

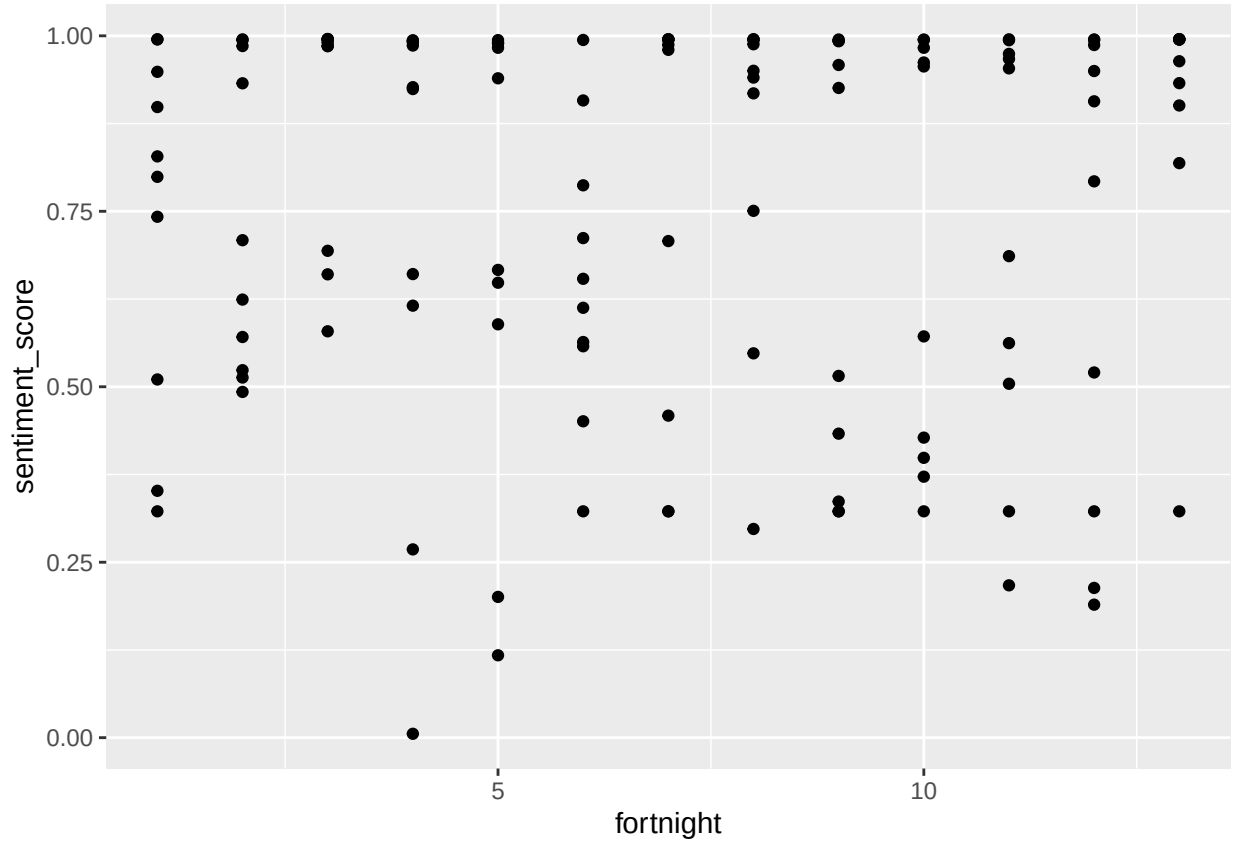
```
p_hat_2_t_pos = 29/50

p_hat_t_pos = (38+29)/(80+50)

sd <- sqrt((((p_hat_t_pos)*(1-p_hat_t_pos))/80)+(((p_hat_t_pos)*(1-p_hat_t_pos))/50))
z_score <- ((p_hat_2_t_pos-p_hat_1_t_pos)-0)/sd

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

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.165) = P(Z <= 1.165) = 0.8781
##  P(X >  1.165) = P(Z >  1.165) = 0.1219
##
```

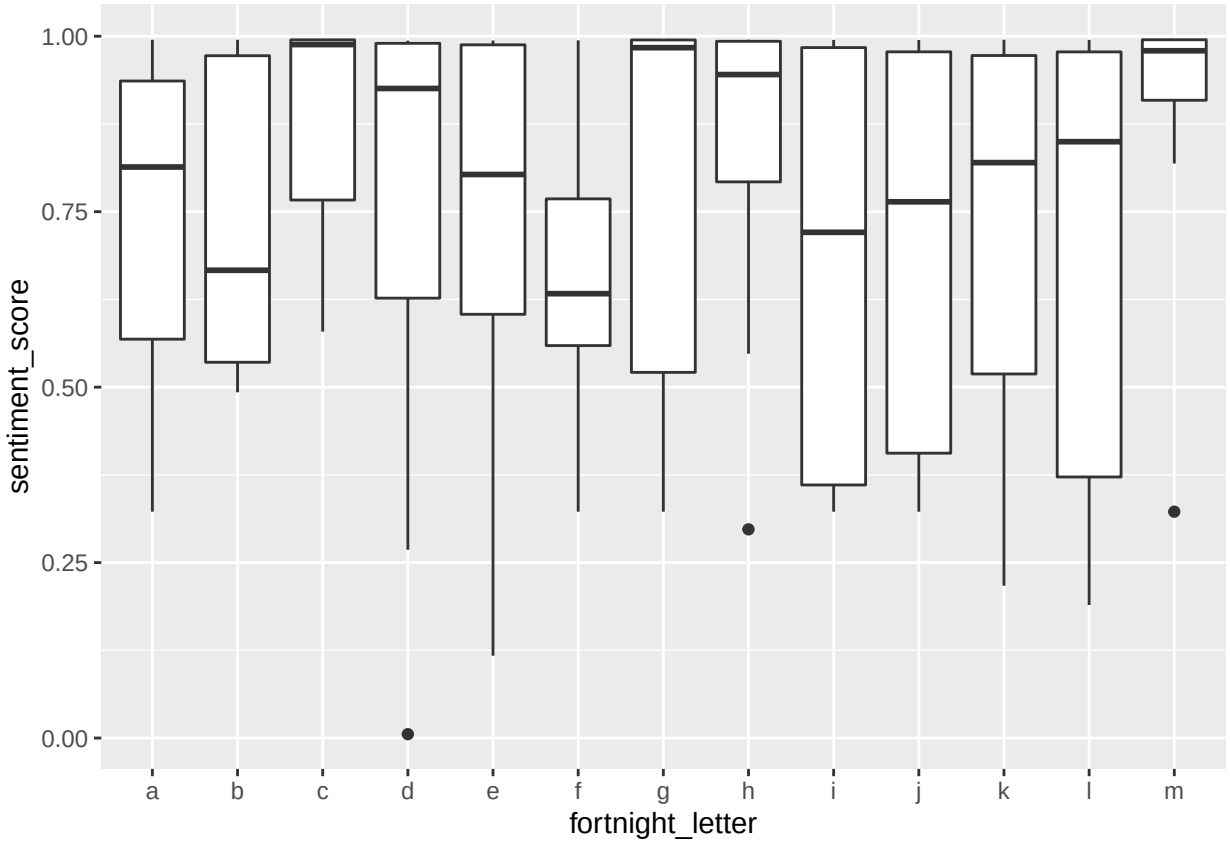


```
## [1] 0.2438481
```

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



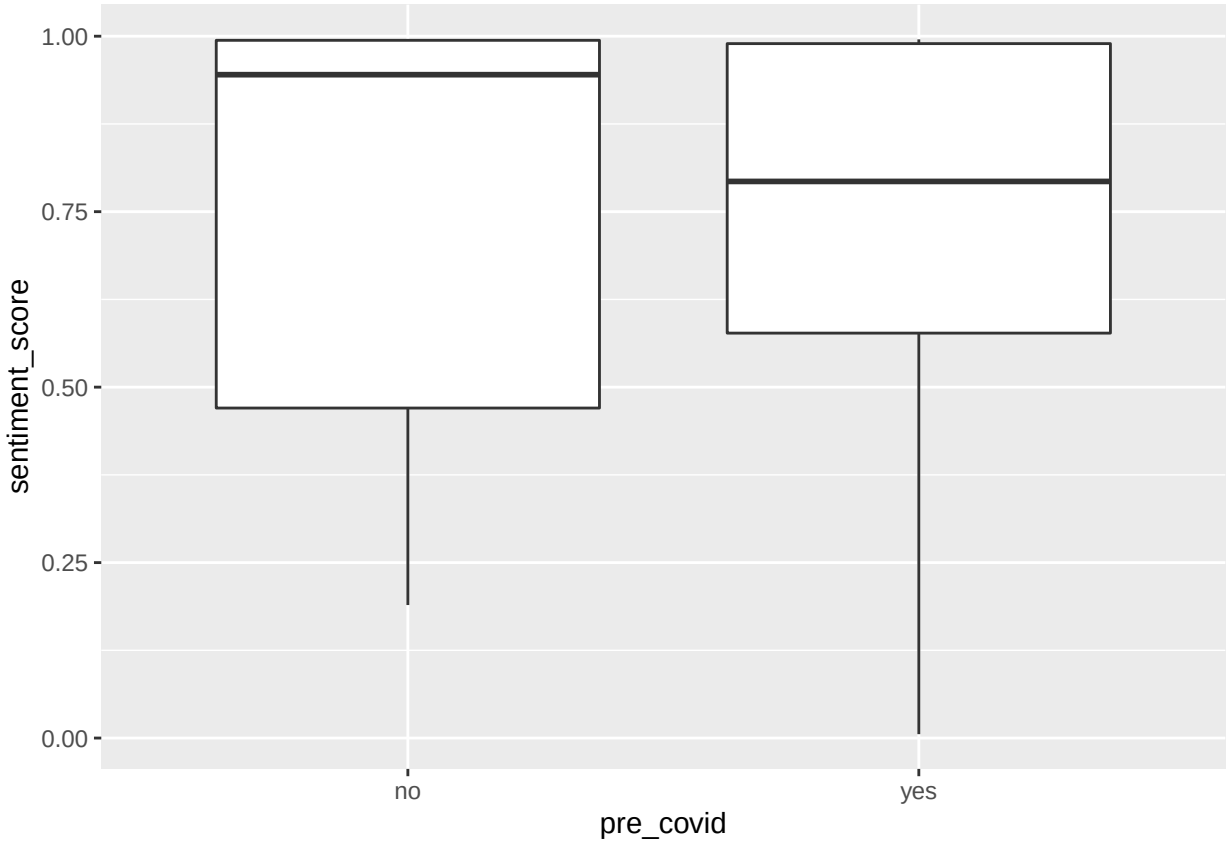

```
ggplot(US_analysis_people) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.739
##  2         2                   0.734
##  3         3                   0.887
##  4         4                   0.736
##  5         5                   0.712
##  6         6                   0.656
##  7         7                   0.776
##  8         8                   0.838
##  9         9                   0.679
## 10        10                   0.698
## 11        11                   0.718
## 12        12                   0.687
## 13        13                   0.891
```

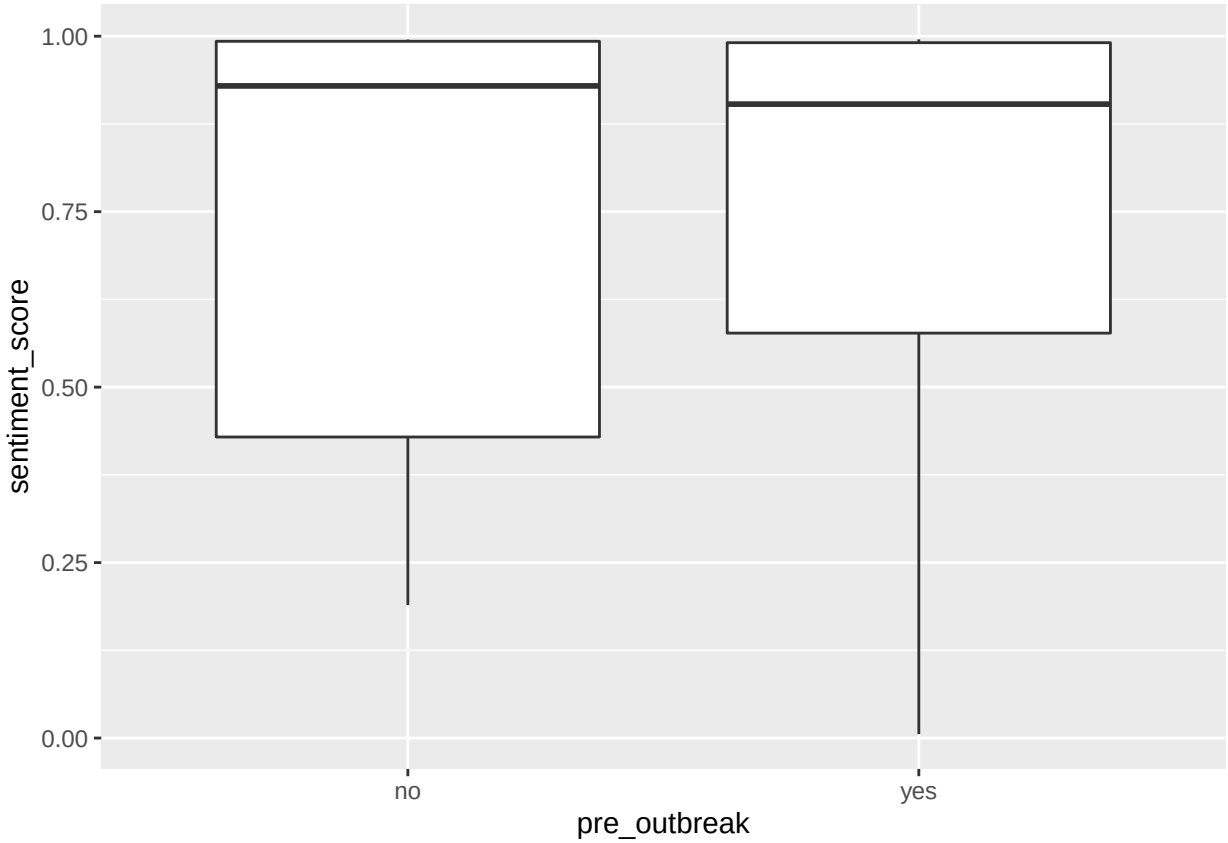
```
ggplot(US_analysis_people) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.755
## 2 yes                         0.744
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.735
## 2 yes                            0.760
```

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

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

```
p_num_precovid = 60
p_num_postcovid = 70
p_num = 130

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

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

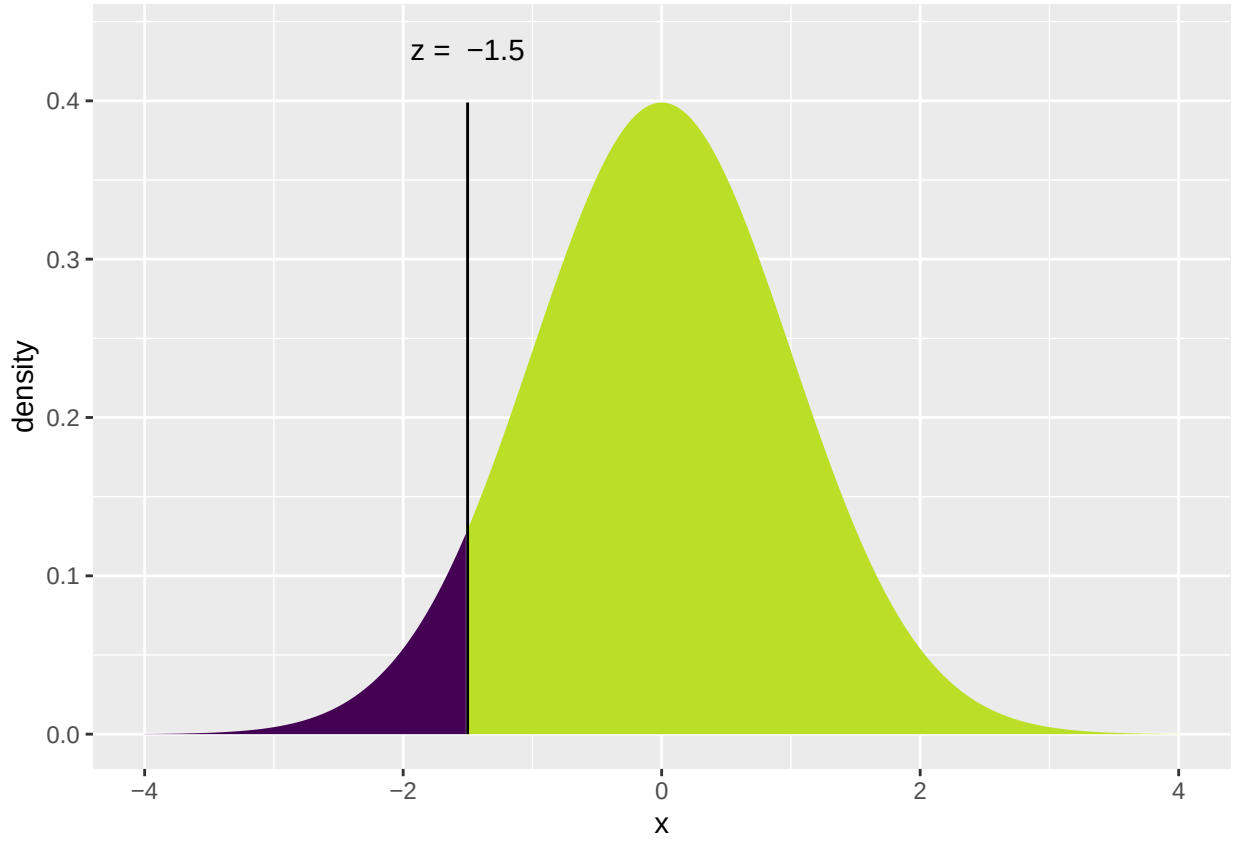
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                       9
## 2 TRUE                       51
```

```
p_hat_1_p_pos = 51/60

US_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                      18
## 2 TRUE                       52
```

```
p_hat_2_p_pos = 52/70

p_hat_p_pos = (51+52)/(60+70)

sd <- sqrt((((p_hat_p_pos)*(1-p_hat_p_pos))/60)+(((p_hat_p_pos)*(1-p_hat_p_pos))/70))
z_score <- ((p_hat_2_p_pos-p_hat_1_p_pos)-0)/sd

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

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.501) = P(Z <= -1.501) = 0.06664
##  P(X >  -1.501) = P(Z >  -1.501) = 0.9334
##
```




```
## [1] 0.1332855
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      50
## 2 TRUE                       80
```

```
p_num_preoutbreak = 80
p_num_postoutbreak = 50
p_num = 130

US_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                      13
## 2 TRUE                       67
```

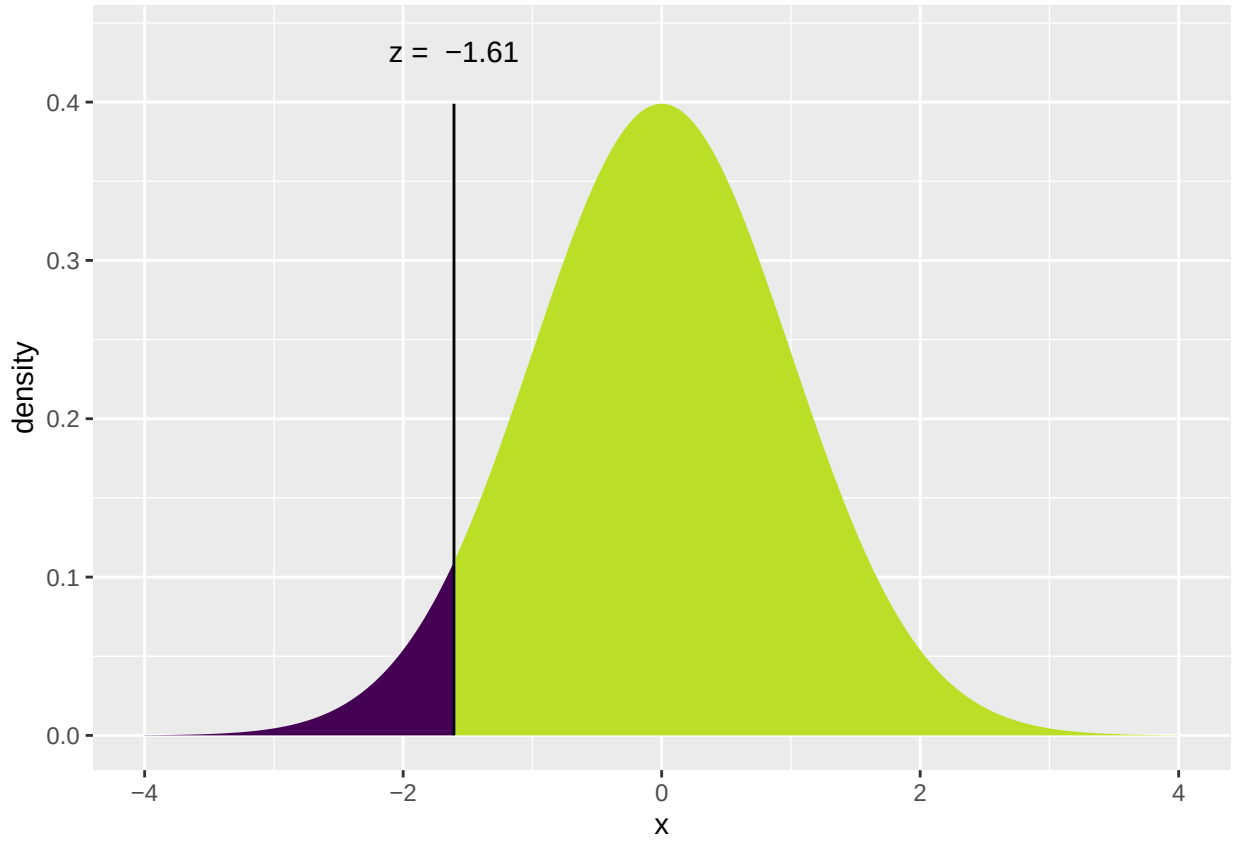
```
p_hat_1_p_pos = 67/80

US_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                      14
## 2 TRUE                       36
```

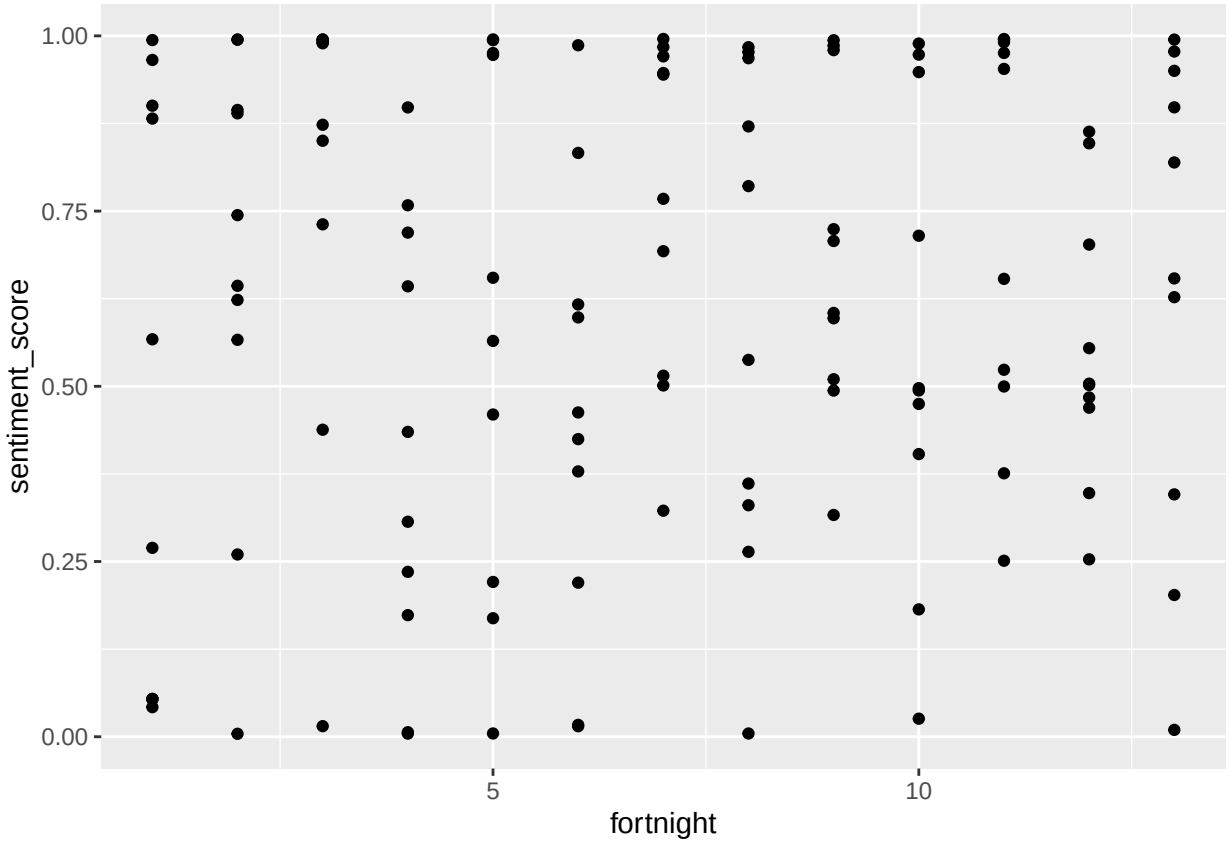
```
p_hat_2_p_pos = 36/50

p_hat_p_pos = (67+36)/(80+50)

sd <- sqrt((((p_hat_p_pos)*(1-p_hat_p_pos))/80)+(((p_hat_p_pos)*(1-p_hat_p_pos))/50))
z_score <- ((p_hat_2_p_pos-p_hat_1_p_pos)-0)/sd

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

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.607) = P(Z <= -1.607) = 0.05406
##  P(X >  -1.607) = P(Z >  -1.607) = 0.9459
##
```

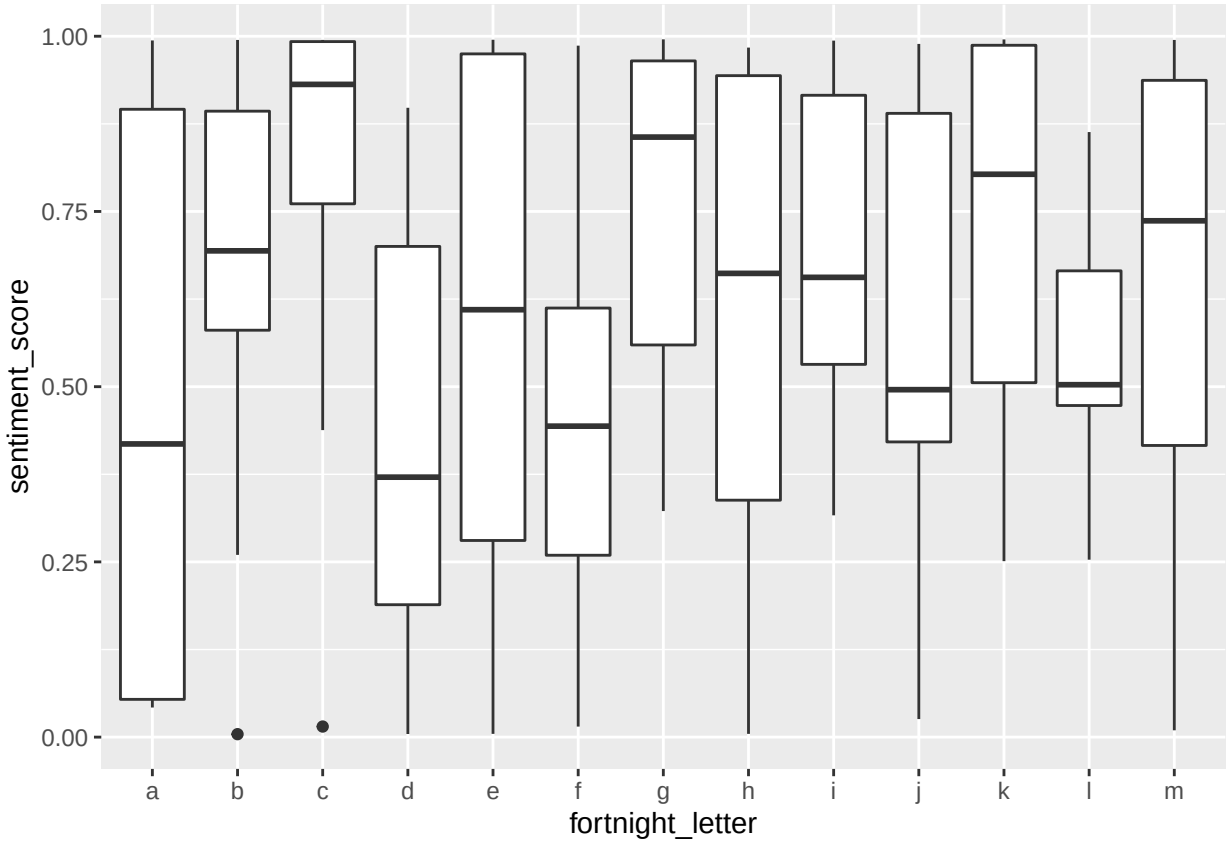


```
## [1] 0.1081167
```

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



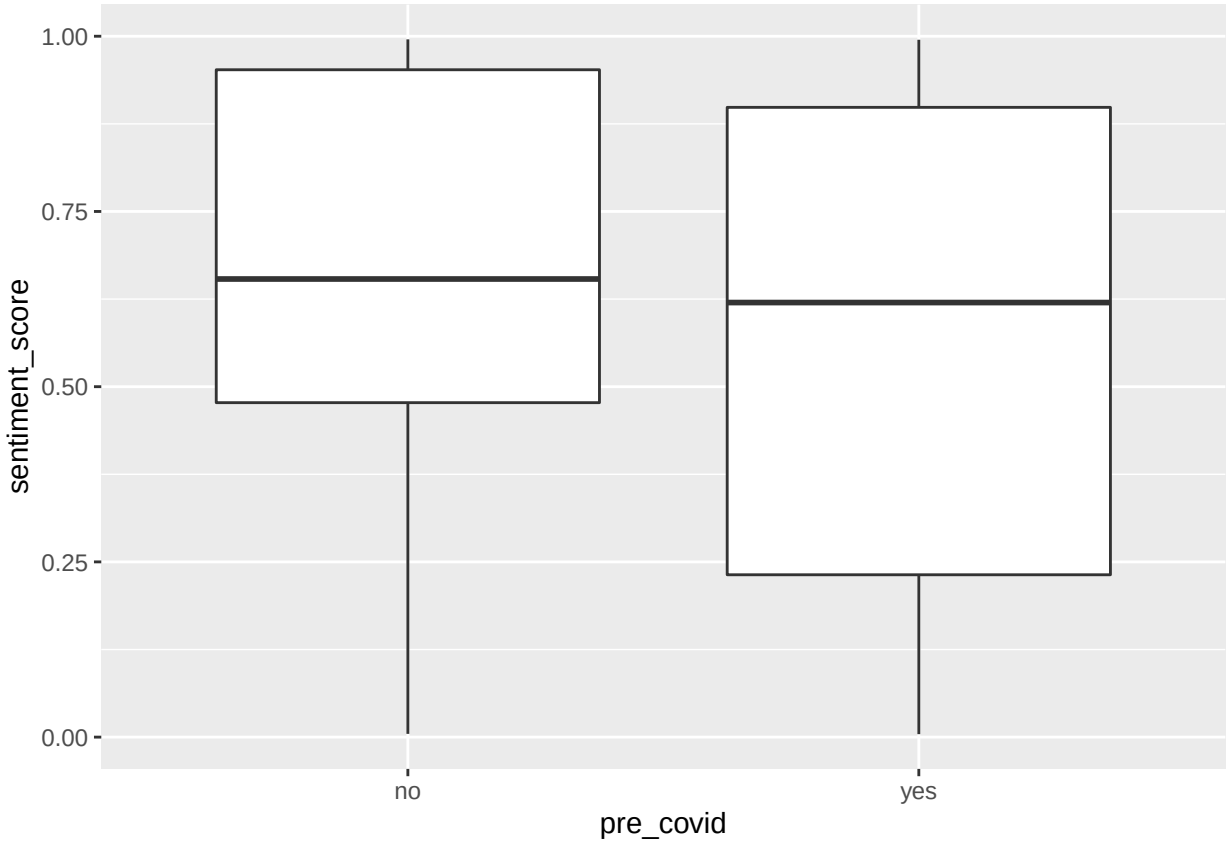

```
ggplot(US_analysis_entertainment) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.478
##  2         2                   0.661
##  3         3                   0.787
##  4         4                   0.418
##  5         5                   0.601
##  6         6                   0.455
##  7         7                   0.764
##  8         8                   0.608
##  9         9                   0.691
## 10        10                   0.570
## 11        11                   0.721
## 12        12                   0.553
## 13        13                   0.648
```

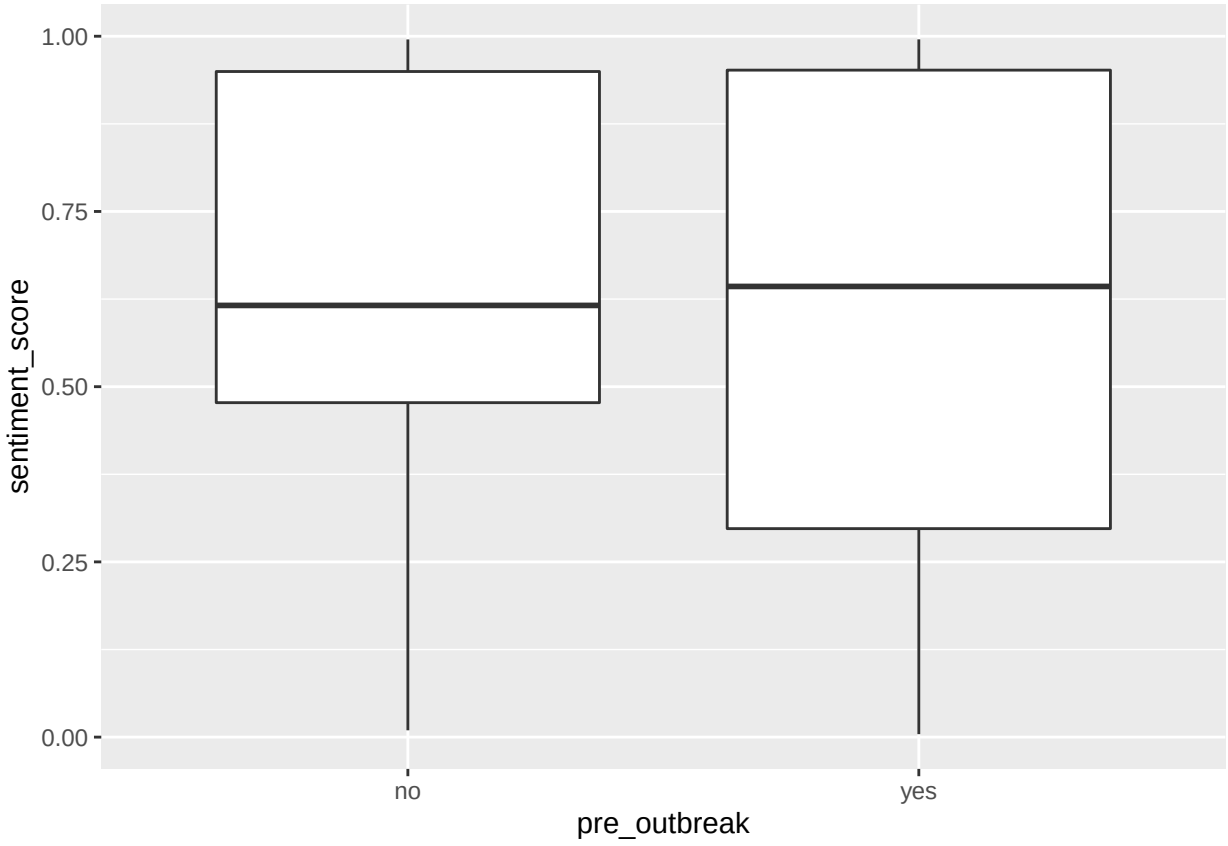
```
ggplot(US_analysis_entertainment) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.651
## 2 yes                         0.567
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.637
## 2 yes                            0.597
```

```
#precovid entertainment
count(US_analysis_entertainment, 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

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

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

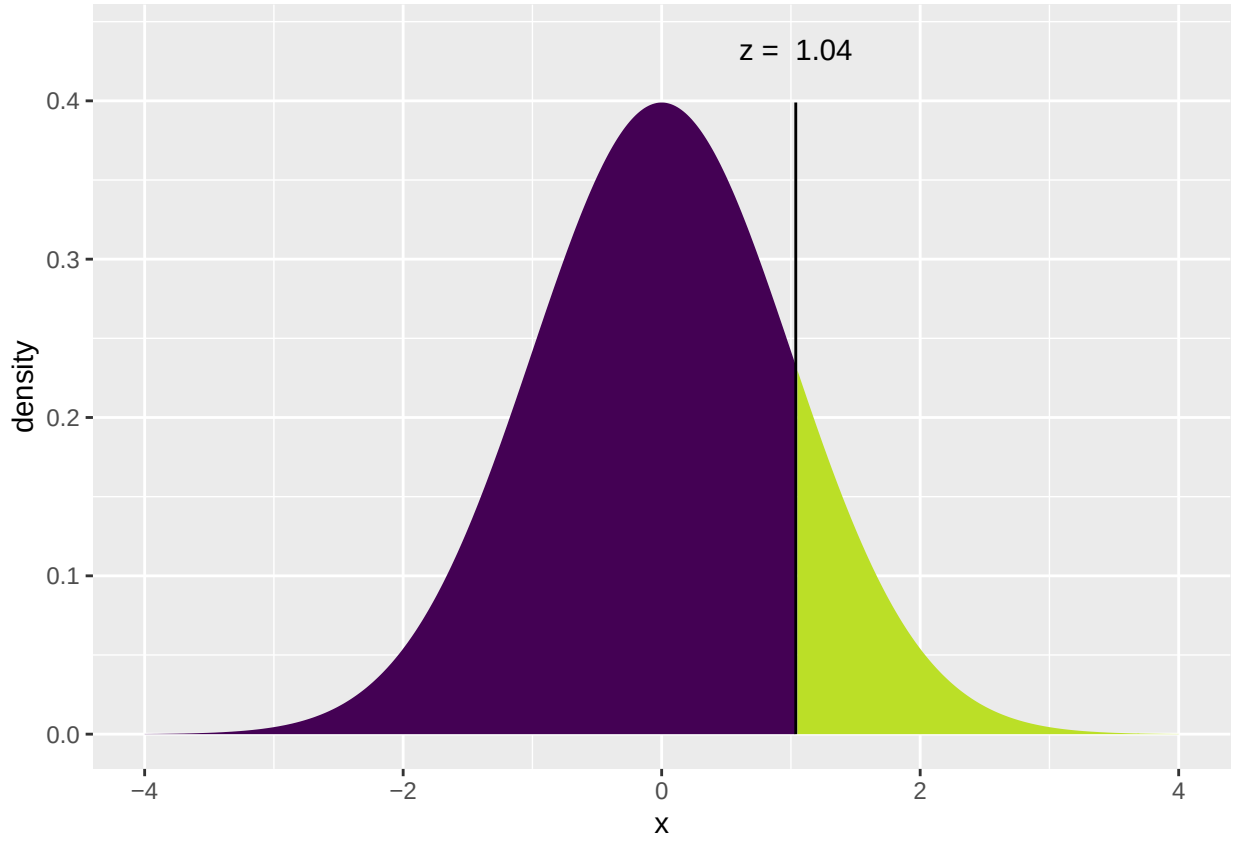
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      25
## 2 TRUE                       35
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 35/60

US_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                      23
## 2 TRUE                       47
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 47/70

p_hat = (35+47)/(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 <= 1.038) = P(Z <= 1.038) = 0.8503
##  P(X >  1.038) = P(Z >  1.038) = 0.1497
##
```




```
## [1] 0.2994661
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      50
## 2 TRUE                       80
```

```
num_preoutbreak = 80
num_postoutbreak = 50
num = 130

US_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                      30
## 2 TRUE                       50
```

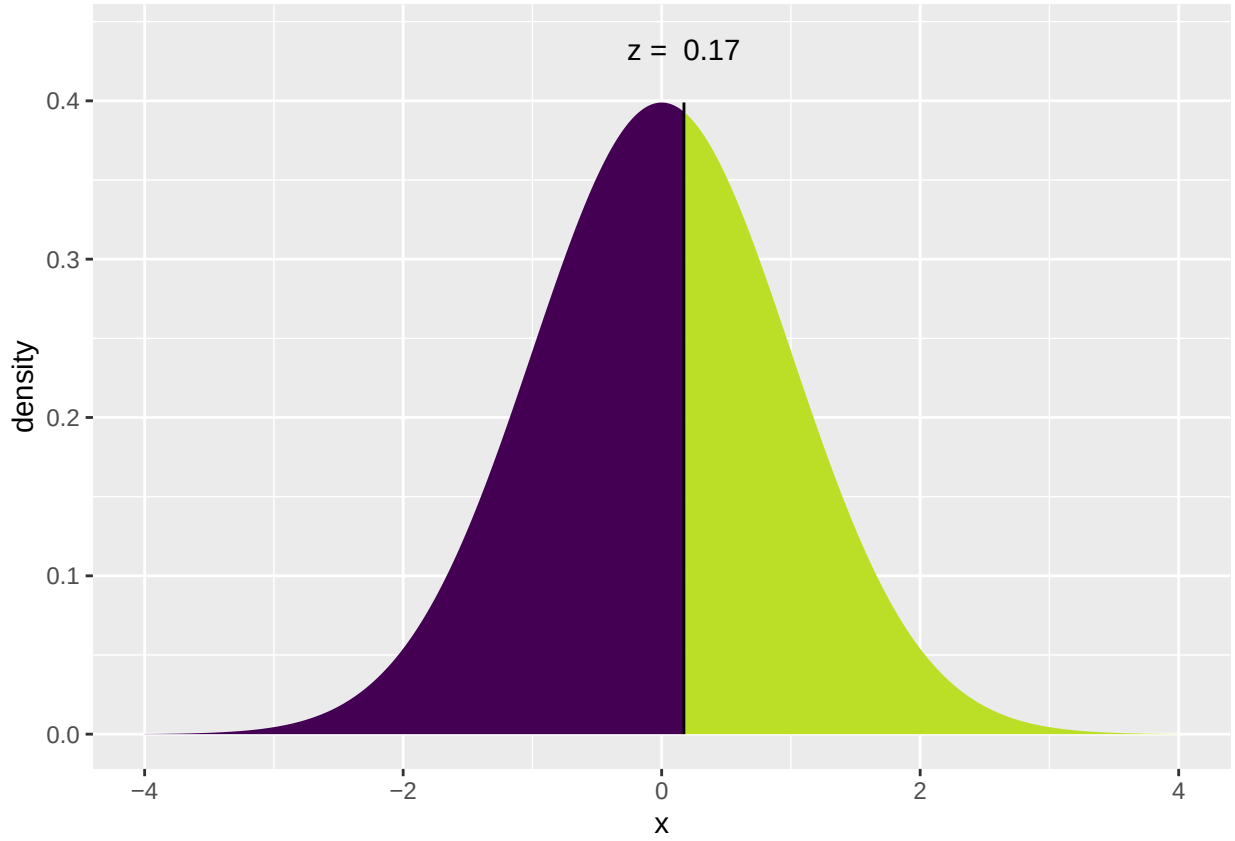
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 50/80
```

```
US_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                      18
## 2 TRUE                       32
```

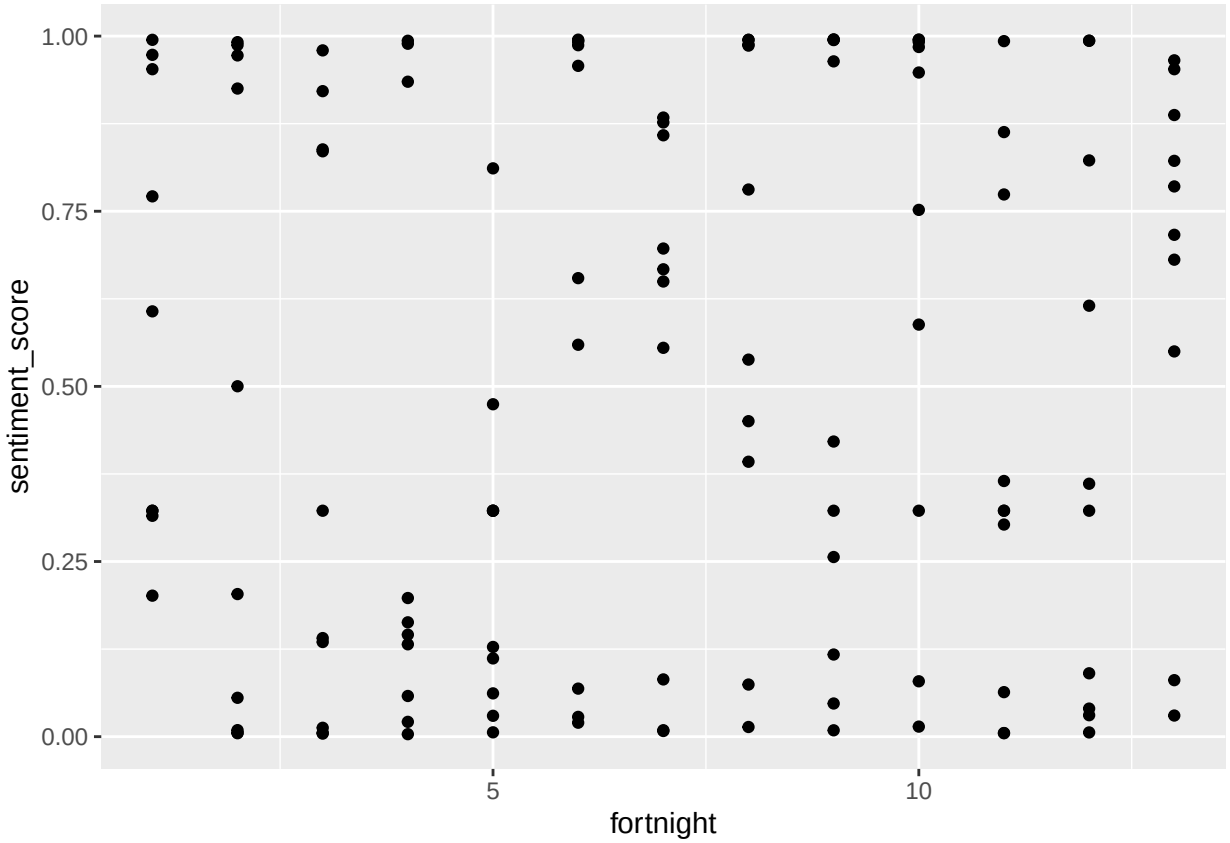
```
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 32/50

p_hat = (50+32)/(80+50)

sd <- sqrt((((p_hat)*(1-p_hat))/80)+(((p_hat)*(1-p_hat))/50))
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 <= 0.1724) = P(Z <= 0.1724) = 0.5684
##  P(X >  0.1724) = P(Z >  0.1724) = 0.4316
##
```

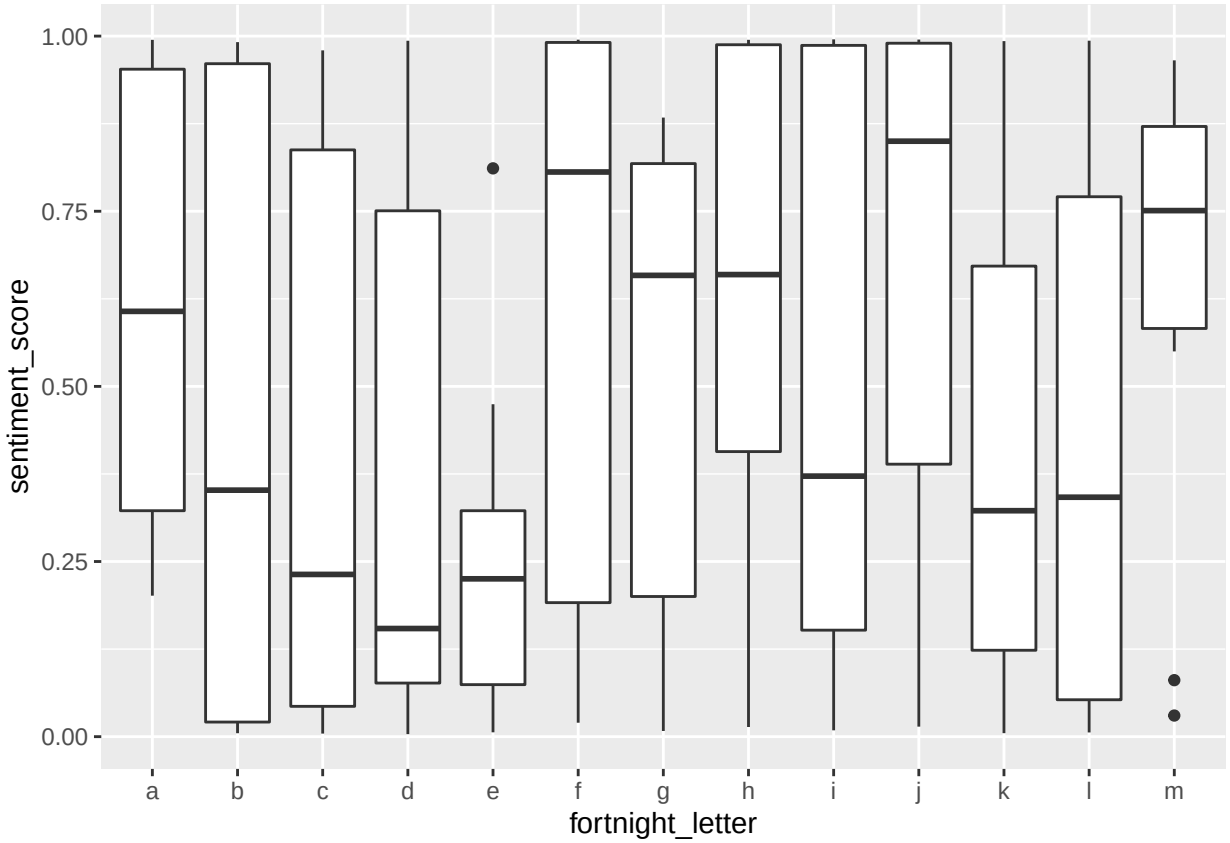



```
## [1] 0.8631143
```

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



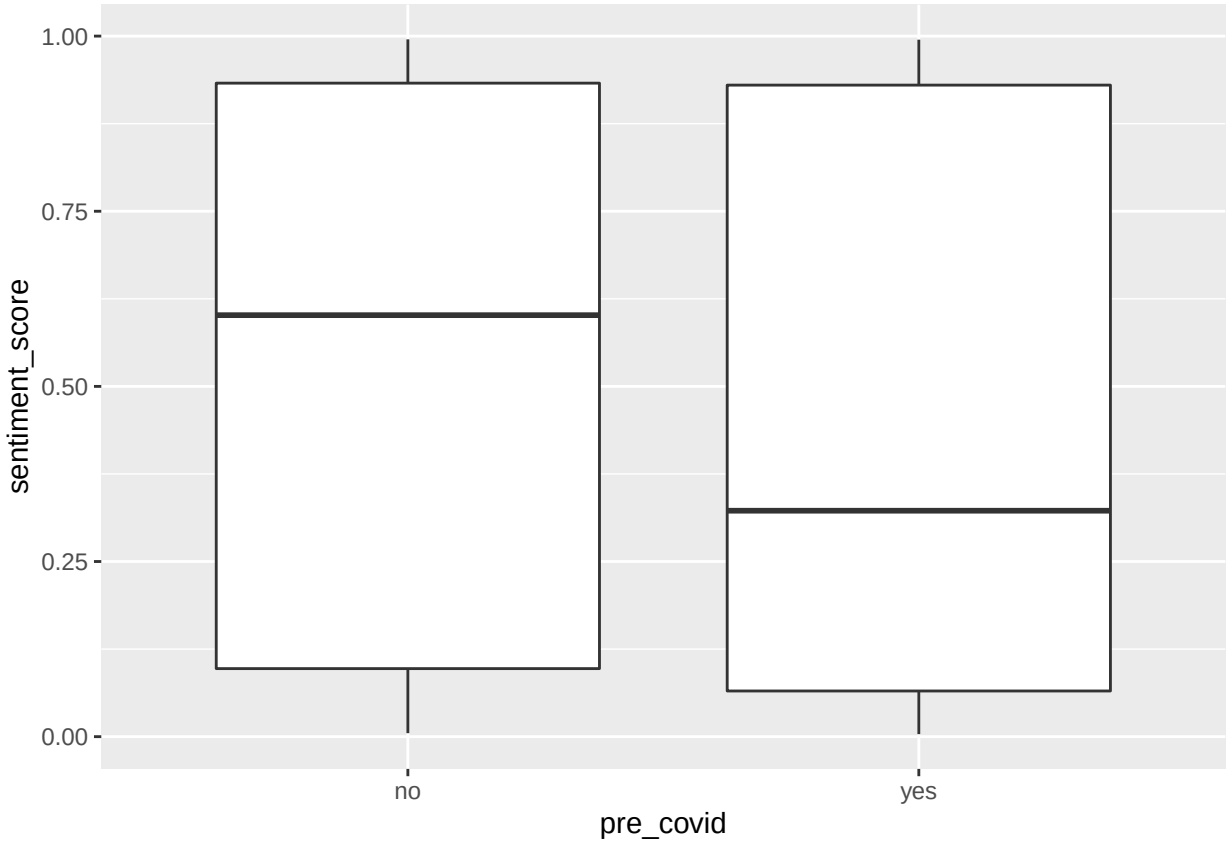

```
ggplot(US_analysis_news) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.607
##  2         2                   0.466
##  3         3                   0.420
##  4         4                   0.364
##  5         5                   0.259
##  6         6                   0.626
##  7         7                   0.529
##  8         8                   0.621
##  9         9                   0.512
## 10        10                   0.667
## 11        11                   0.402
## 12        12                   0.428
## 13        13                   0.647
```

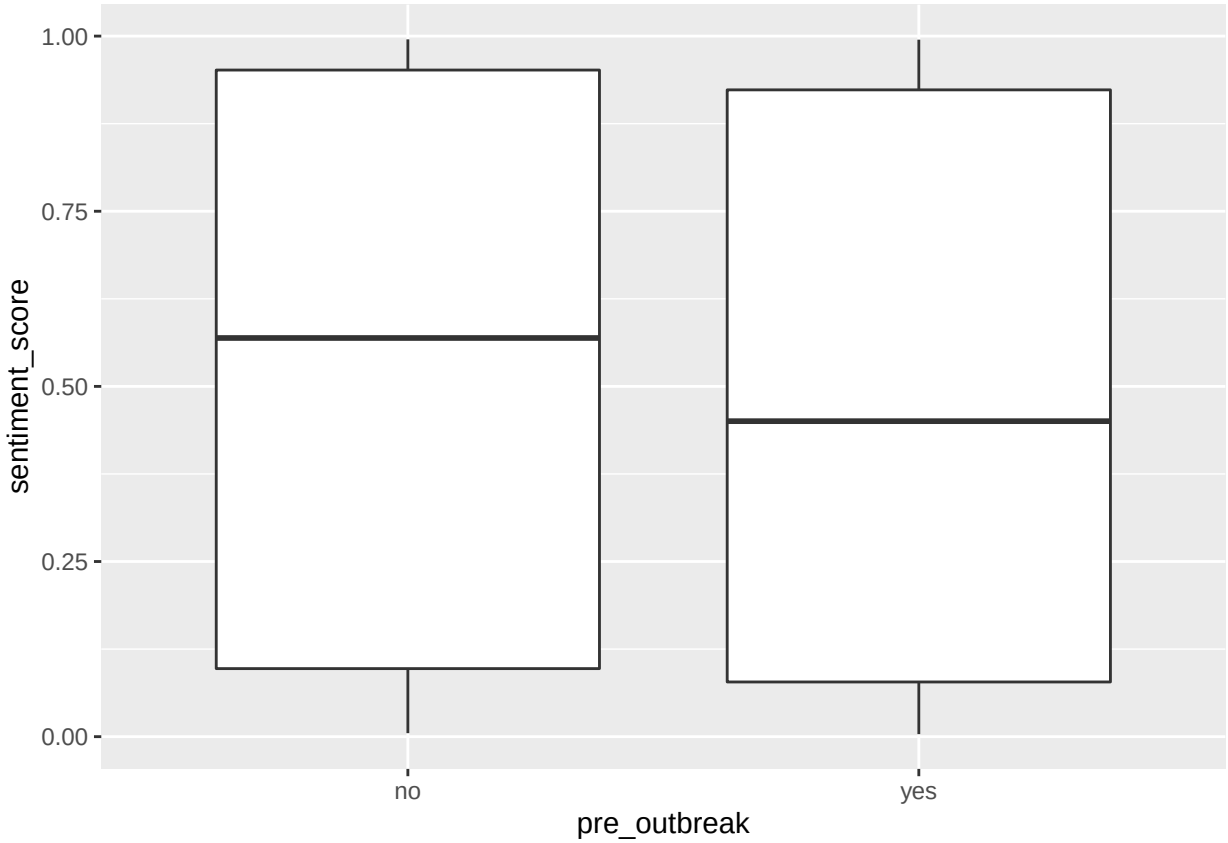
```
ggplot(US_analysis_news) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.544
## 2 yes                         0.454
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.531
## 2 yes                            0.485
```

```
#precovid news
count(US_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

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

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

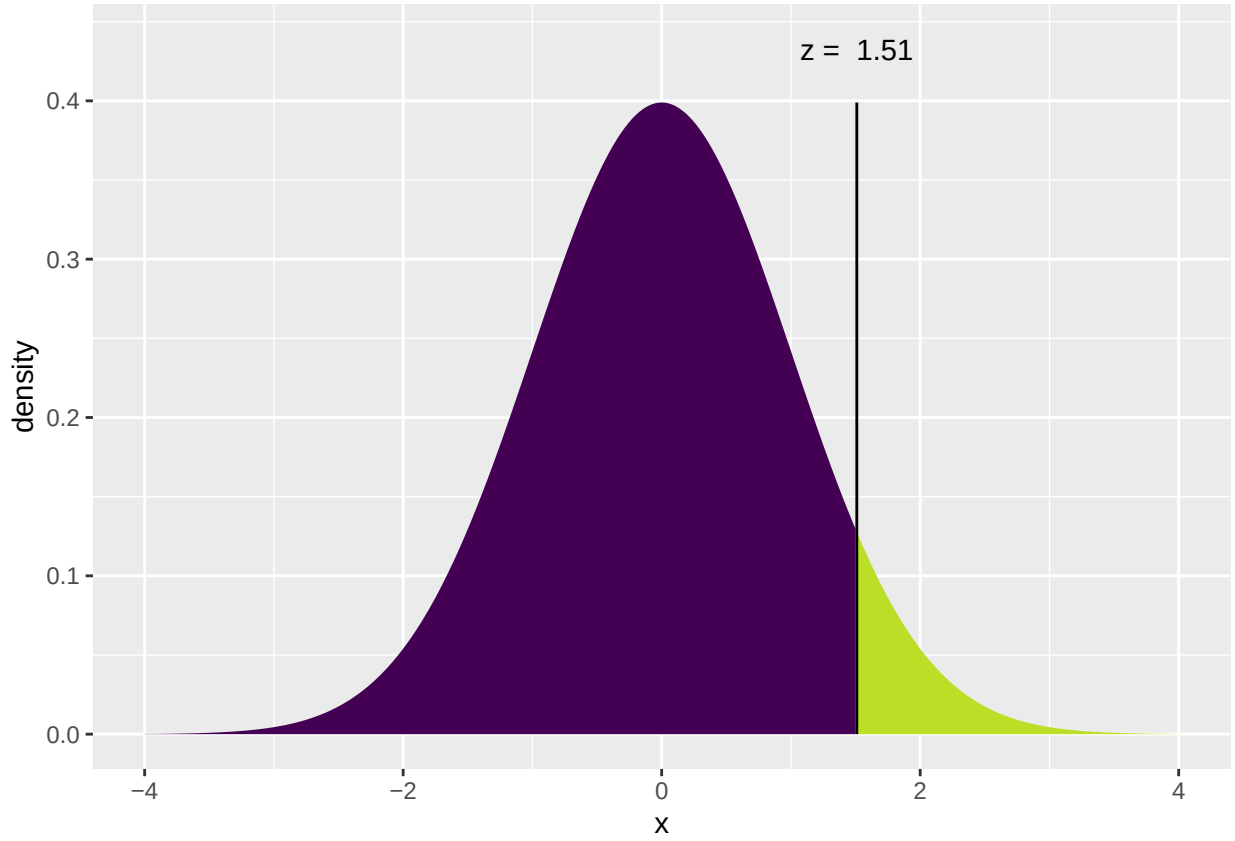
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      34
## 2 TRUE                       25
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 25/59

US_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                      31
## 2 TRUE                       39
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 39/70

p_hat = (25+39)/(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* (1-xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.51) = P(Z <= 1.51) = 0.9345
##  P(X >  1.51) = P(Z >  1.51) = 0.06554
##
```




```
## [1] 0.1310897
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      50
## 2 TRUE                       79
```

```
num_preoutbreak = 79
num_postoutbreak = 50
num = 129

US_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                      41
## 2 TRUE                       38
```

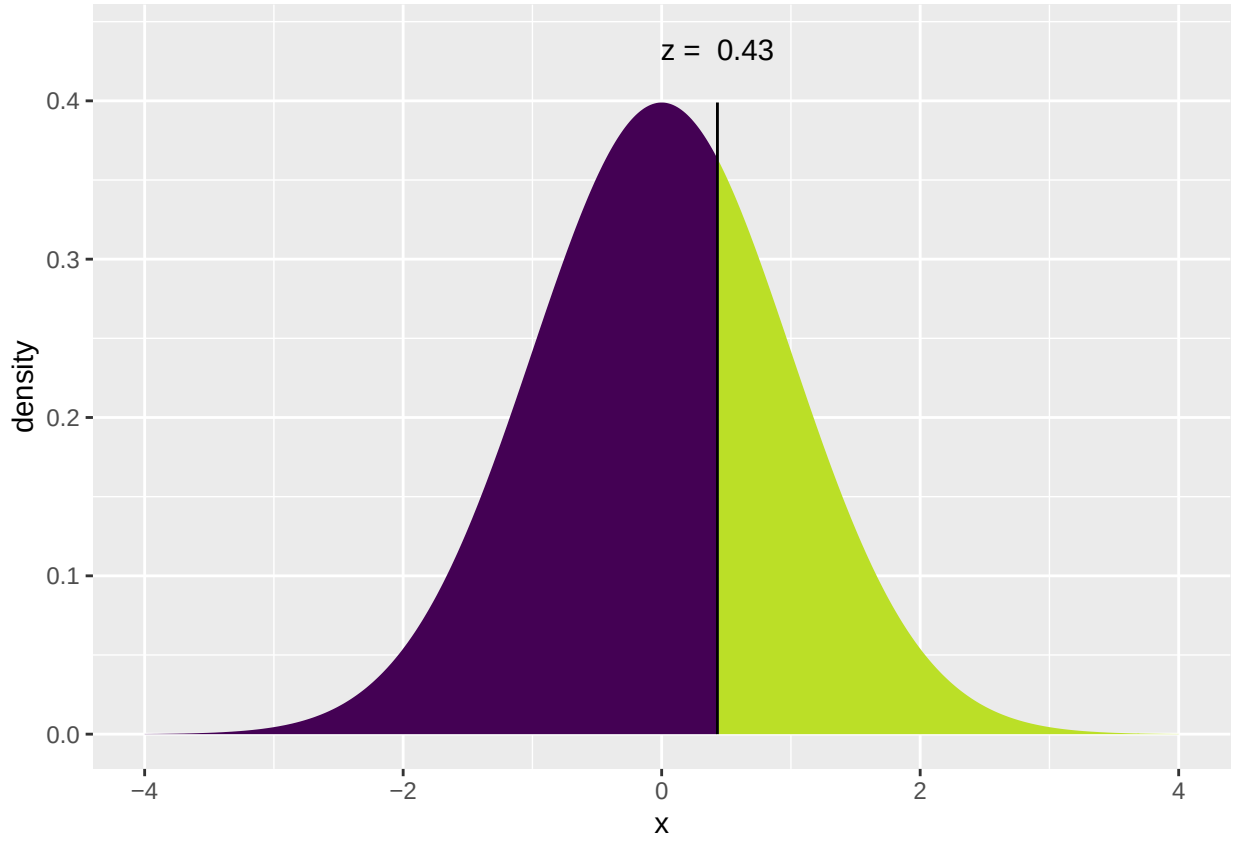
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 38/79
```

```r
US_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                      24
## 2 TRUE                       26
```

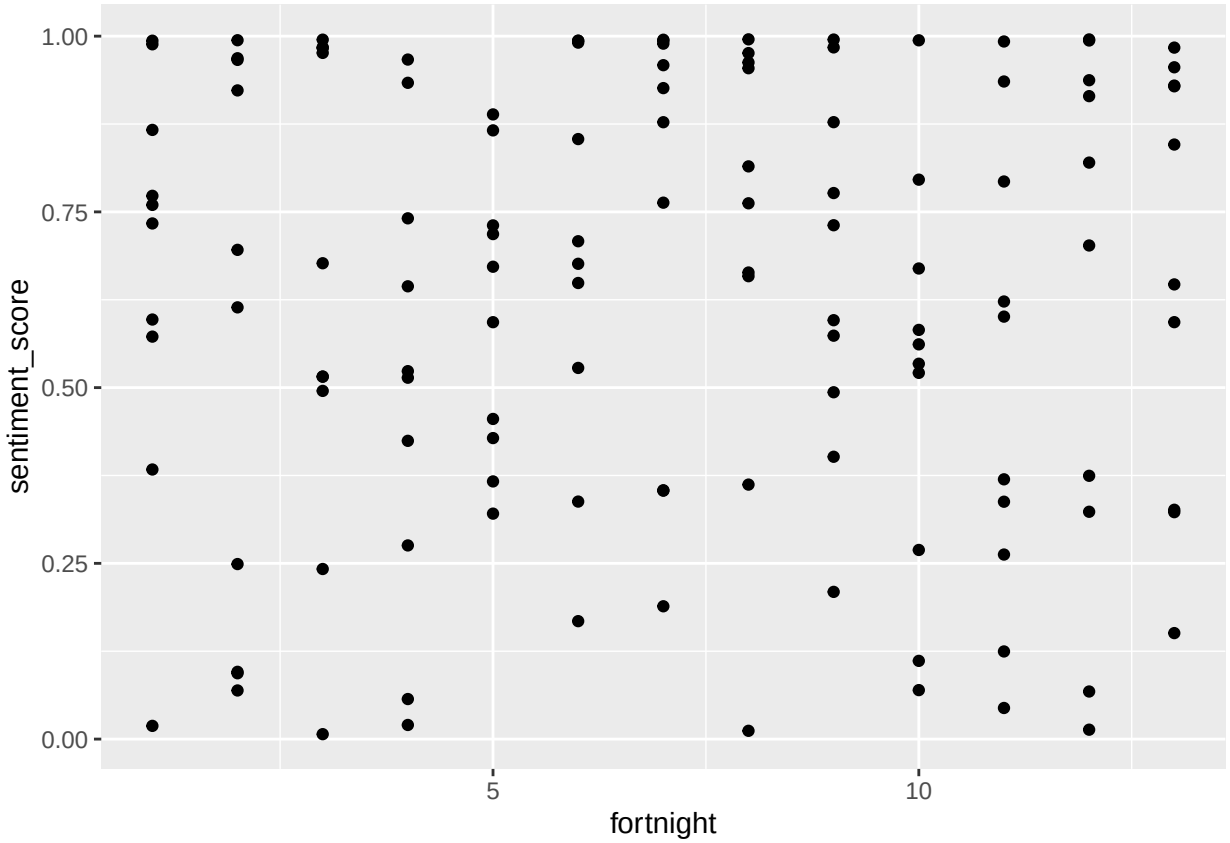
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 26/50

p_hat = (38+26)/(79+50)

sd <- sqrt((((p_hat)*(1-p_hat))/79)+(((p_hat)*(1-p_hat))/50))
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 <= 0.4315) = P(Z <= 0.4315) = 0.6669
##  P(X >  0.4315) = P(Z >  0.4315) = 0.3331
##
```

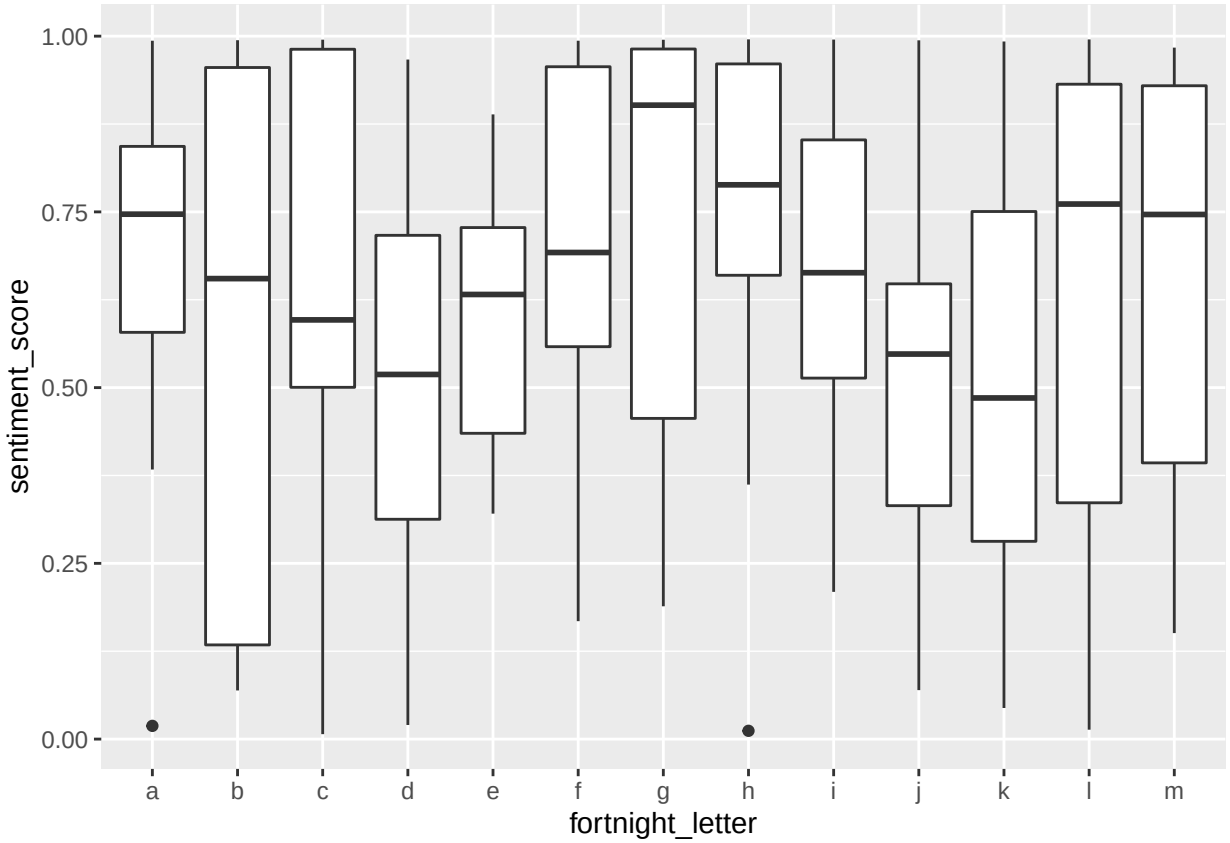



```
## [1] 0.6661124
```

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



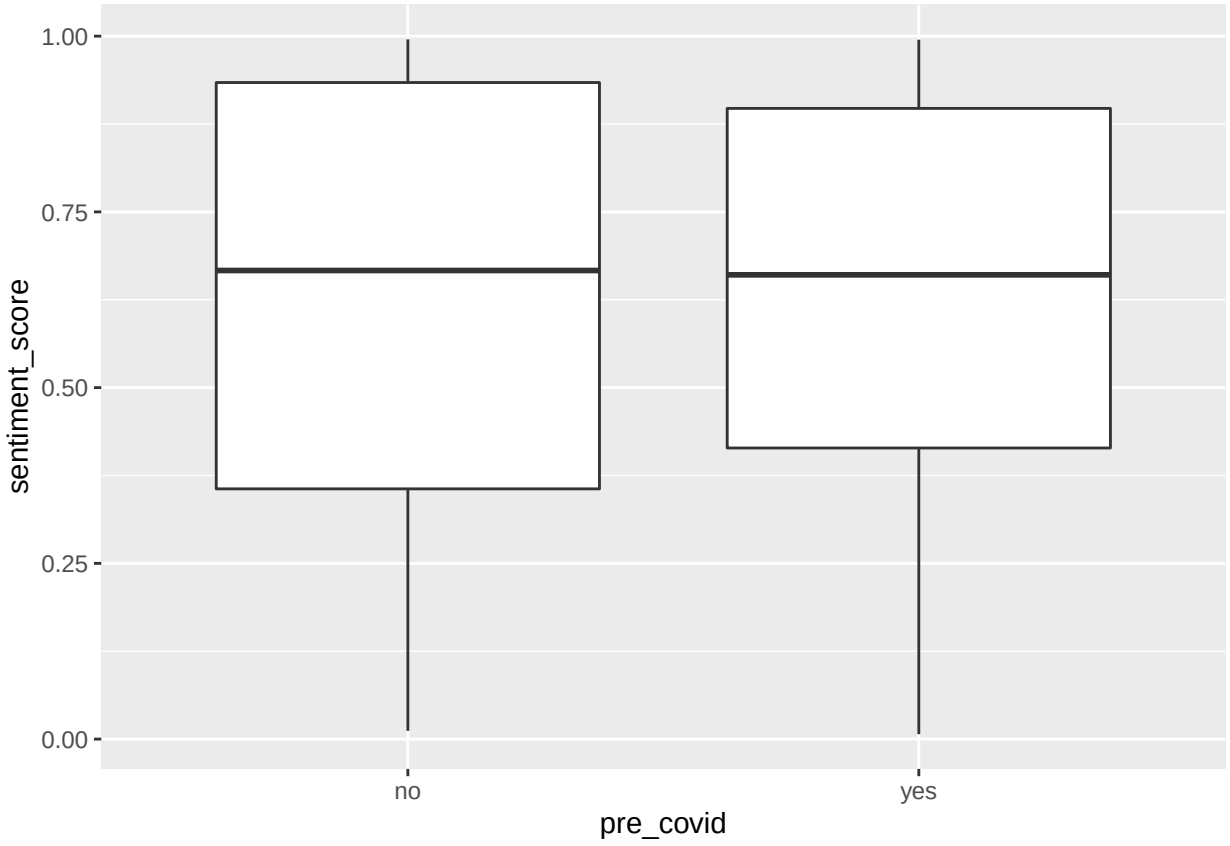

```
ggplot(US_analysis_how_to) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.669
##  2         2                   0.567
##  3         3                   0.639
##  4         4                   0.510
##  5         5                   0.604
##  6         6                   0.690
##  7         7                   0.740
##  8         8                   0.716
##  9         9                   0.664
## 10        10                   0.511
## 11        11                   0.508
## 12        12                   0.614
## 13        13                   0.668
```

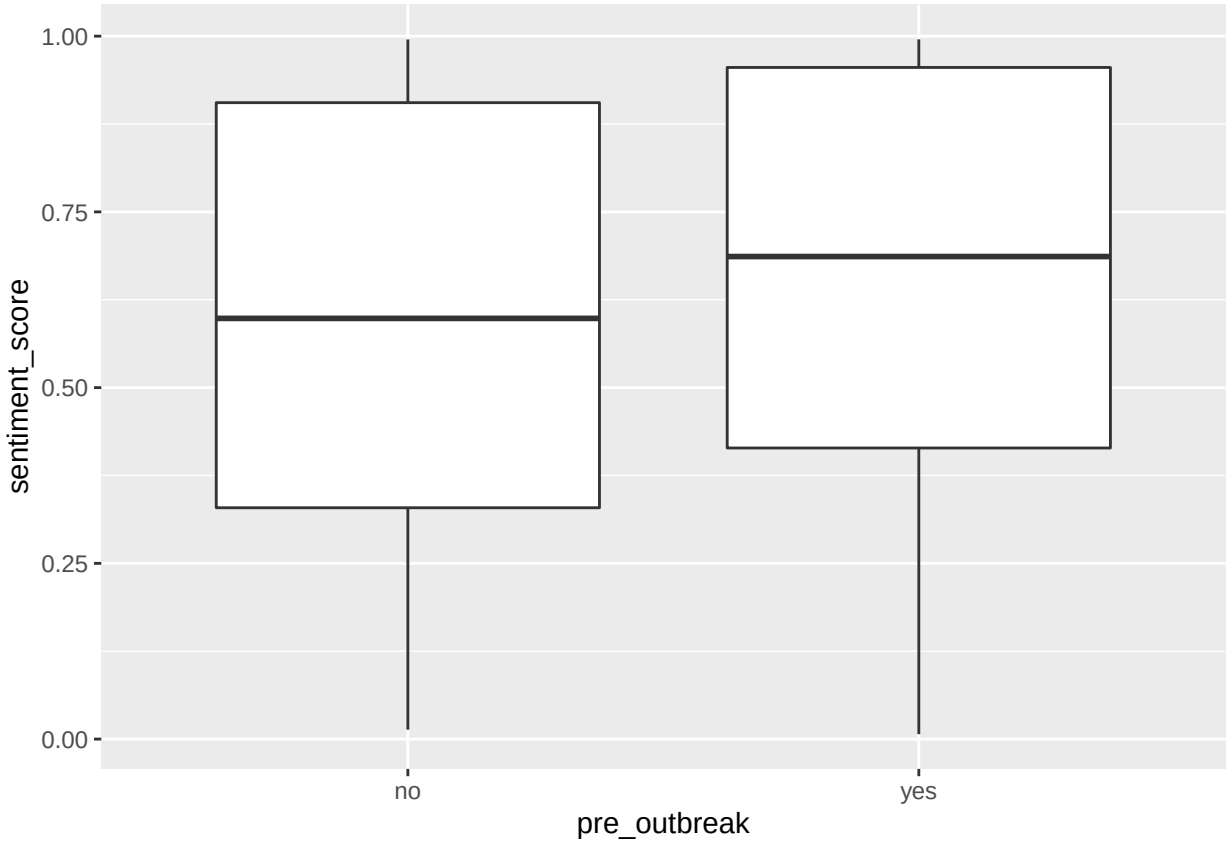
```
ggplot(US_analysis_how_to) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




```
US_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                          0.632
## 2 yes                         0.613
```

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




```
US_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                             0.593
## 2 yes                            0.642
```

```
#precovid how-to
count(US_analysis_how_to, 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

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

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

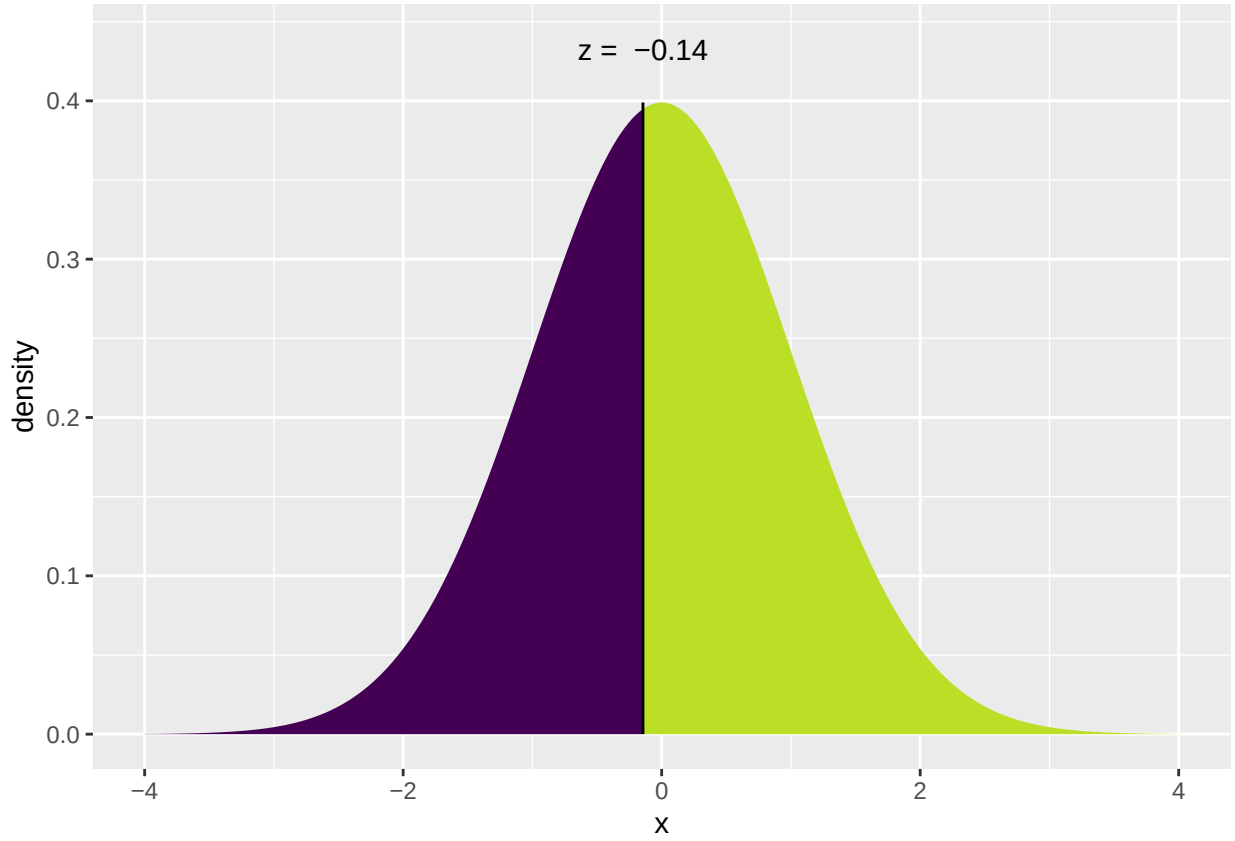
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      19
## 2 TRUE                       41
```

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

US_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                      23
## 2 TRUE                       47
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 47/70

p_hat = (41+47)/(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* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.1447) = P(Z <= -0.1447) = 0.4425
##  P(X >  -0.1447) = P(Z >  -0.1447) = 0.5575
##
```




```
## [1] 0.8849523
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      50
## 2 TRUE                       80
```

```
num_preoutbreak = 80
num_postoutbreak = 50
num = 130

US_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                      24
## 2 TRUE                       56
```

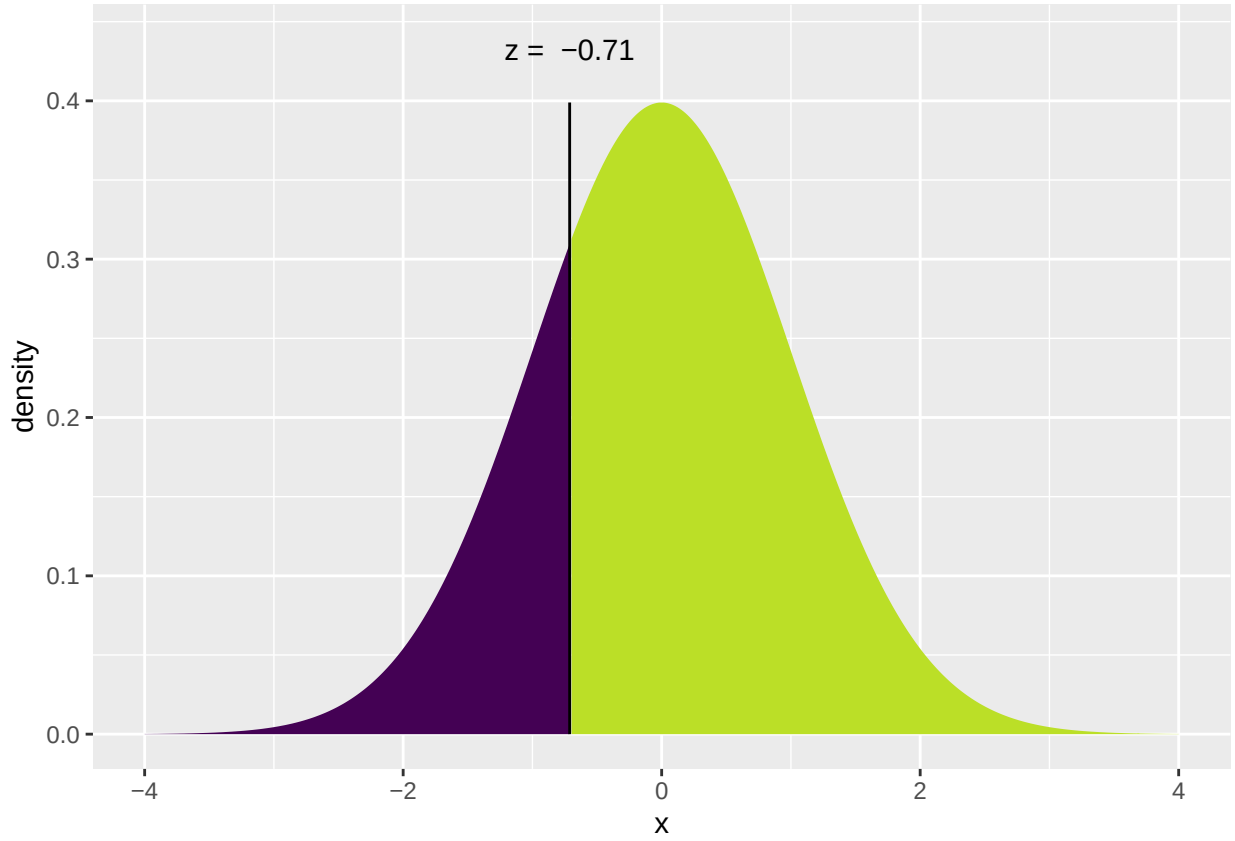
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 56/80
```

```r
US_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                       32
```

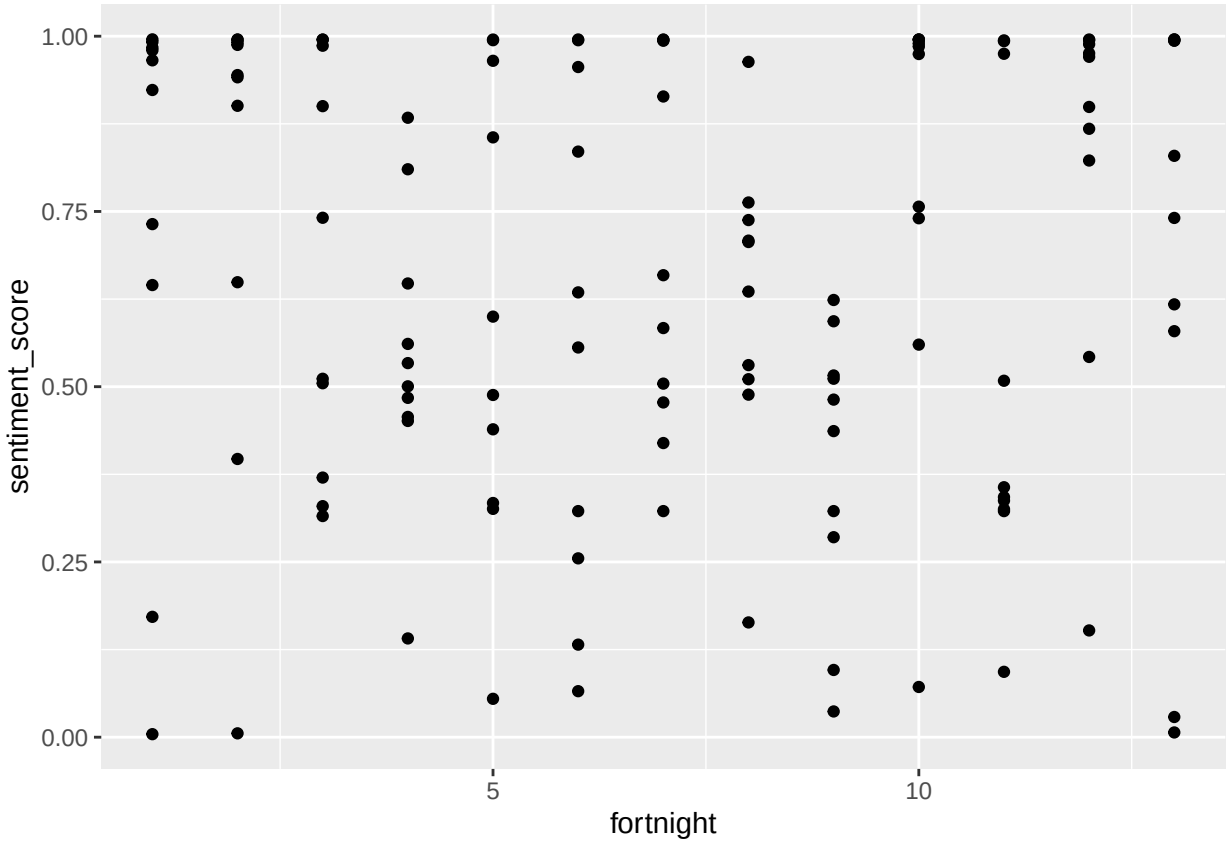
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 32/50

p_hat = (56+32)/(80+50)

sd <- sqrt((((p_hat)*(1-p_hat))/80)+(((p_hat)*(1-p_hat))/50))
z_score <- ((p_hat2-p_hat1)-0)/sd

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

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.7117) = P(Z <= -0.7117) = 0.2383
##  P(X >  -0.7117) = P(Z >  -0.7117) = 0.7617
##
```

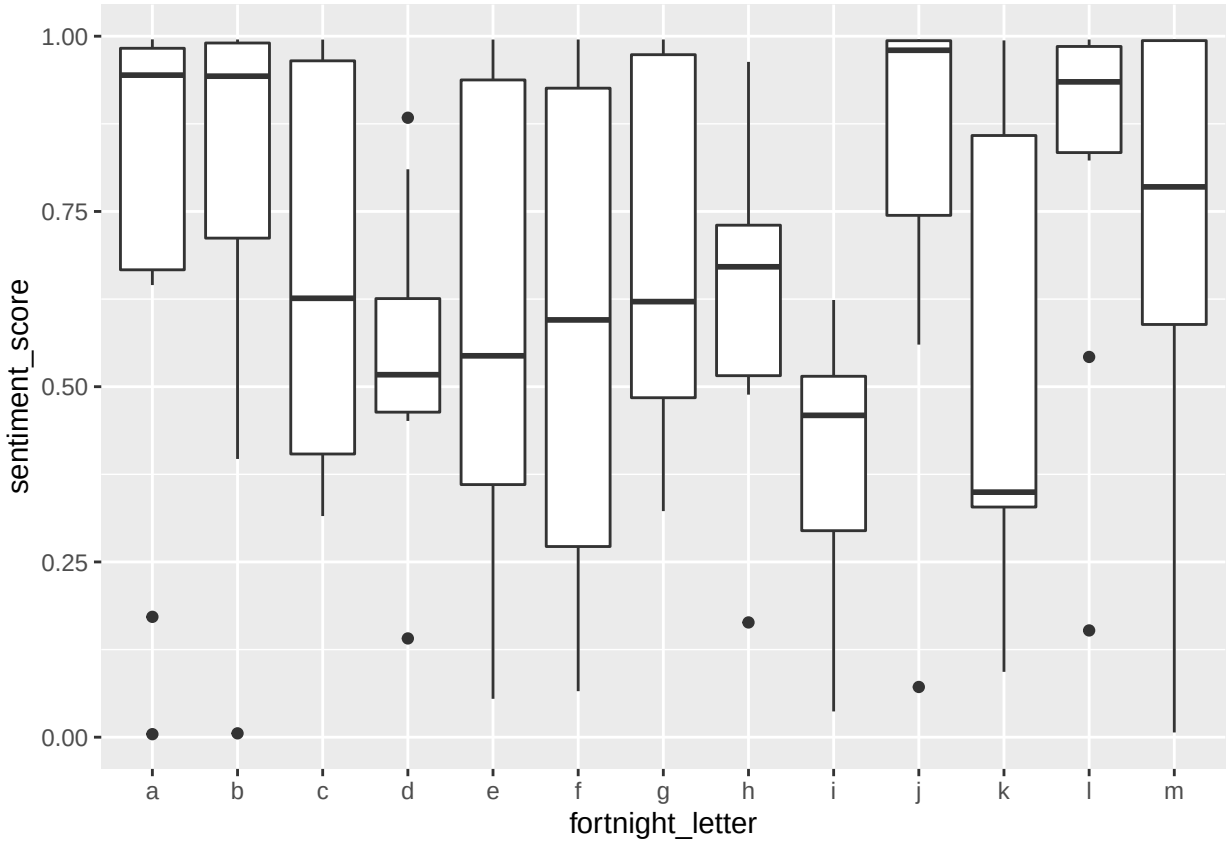



```
## [1] 0.4766607
```

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



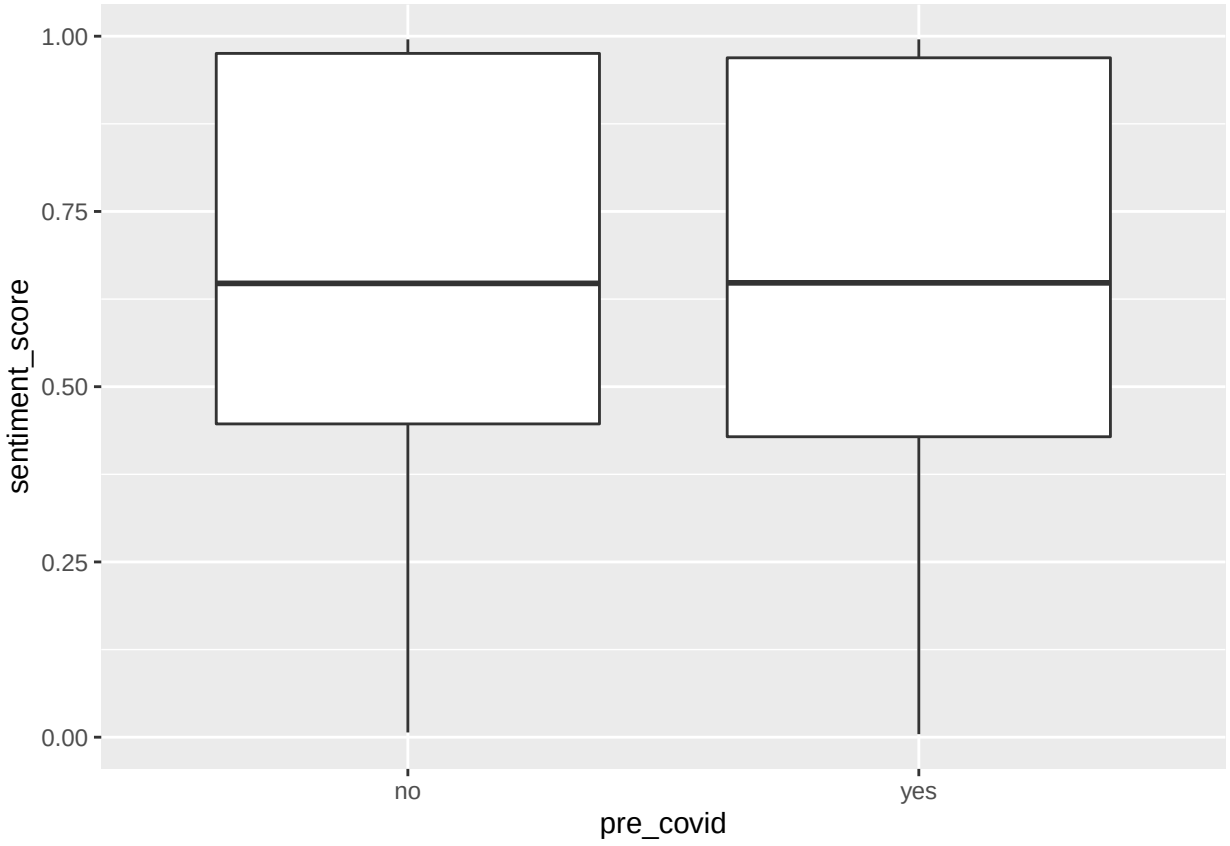

```
ggplot(US_analysis_education) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.739
##  2         2                   0.781
##  3         3                   0.665
##  4         4                   0.547
##  5         5                   0.605
##  6         6                   0.575
##  7         7                   0.686
##  8         8                   0.621
##  9         9                   0.390
## 10        10                   0.806
## 11        11                   0.525
## 12        12                   0.821
## 13        13                   0.678
```

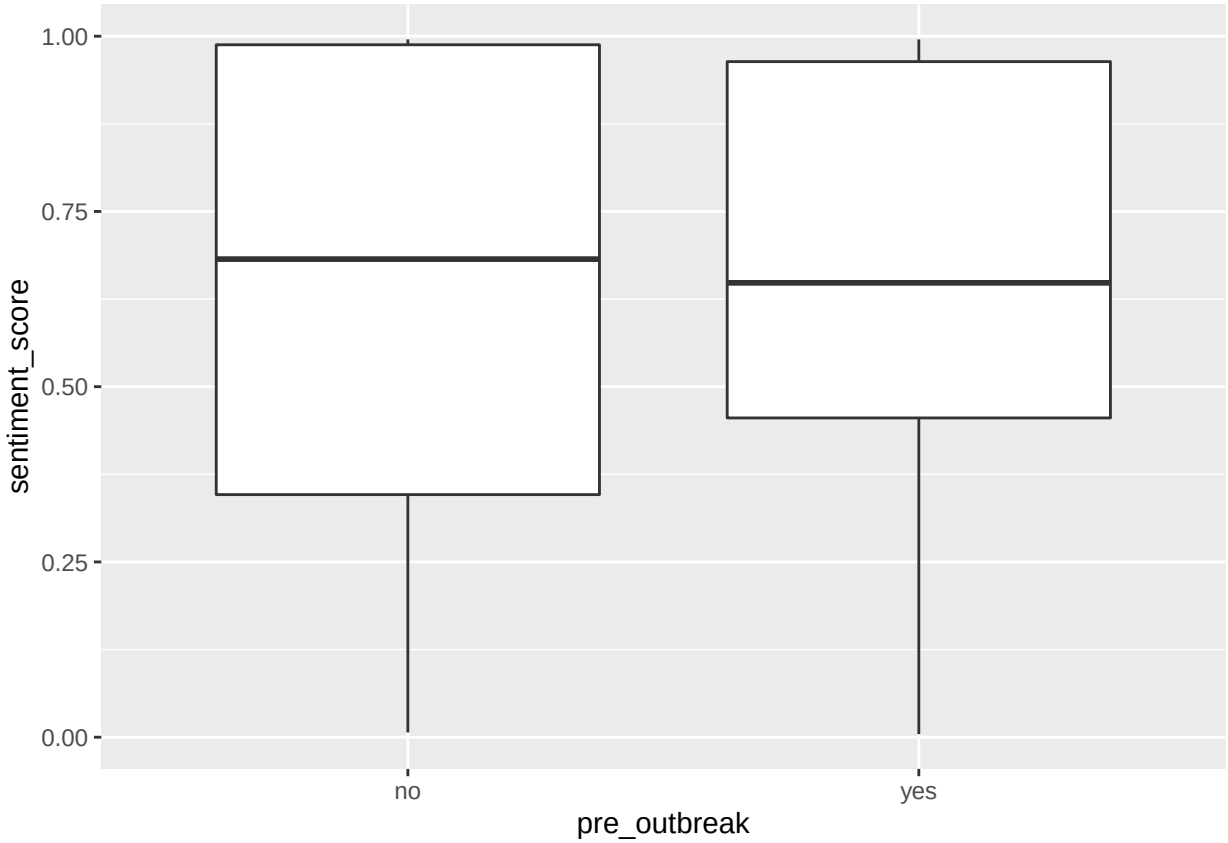
```
ggplot(US_analysis_education) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.647
## 2 yes                         0.652
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.644
## 2 yes                            0.652
```

```
#precovid education
count(US_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

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

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

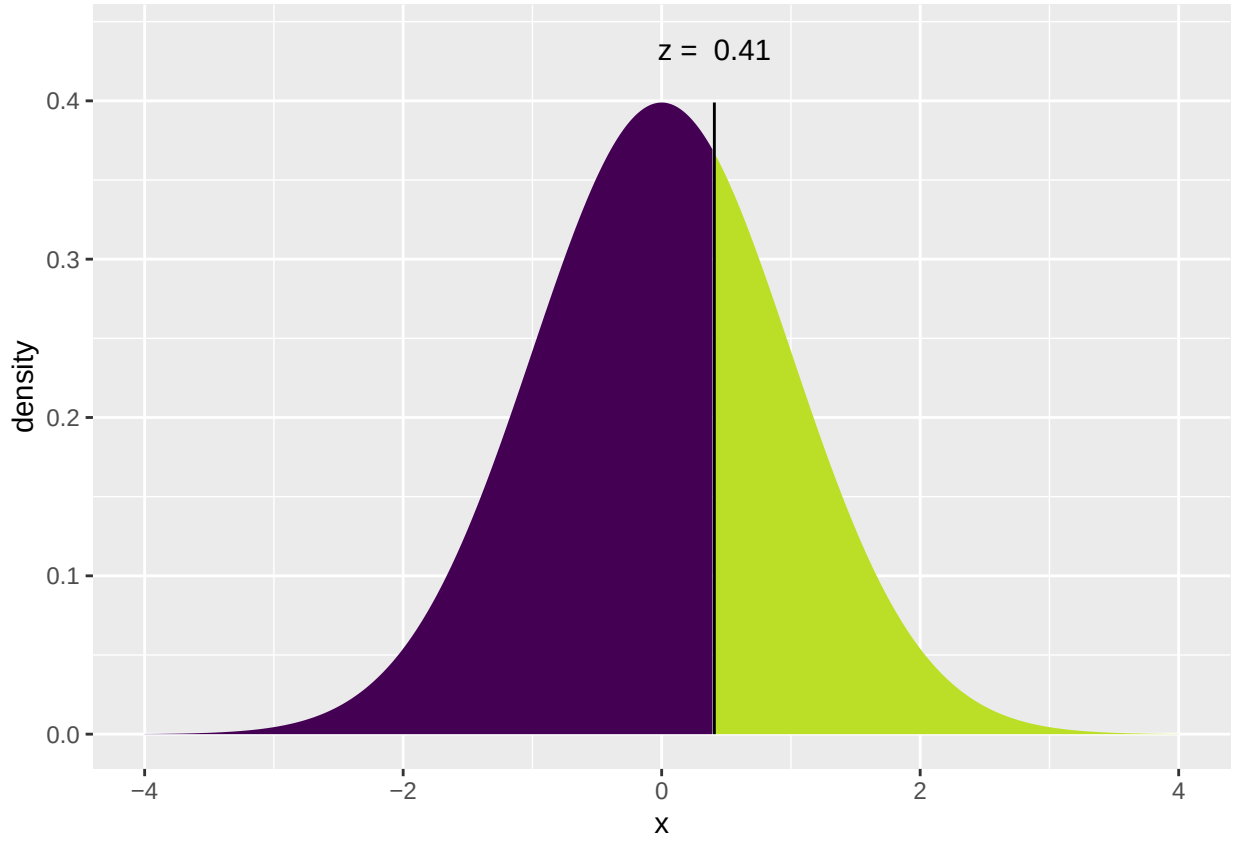
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      20
## 2 TRUE                       40
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 40/60

US_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                      21
## 2 TRUE                       49
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 = 49/70

p_hat = (40+49)/(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 <= 0.4077) = P(Z <= 0.4077) = 0.6583
##  P(X >  0.4077) = P(Z >  0.4077) = 0.3417
##
```




```
## [1] 0.6834612
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      50
## 2 TRUE                       80
```

```
num_preoutbreak = 80
num_postoutbreak = 50
num = 130

US_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                      25
## 2 TRUE                       55
```

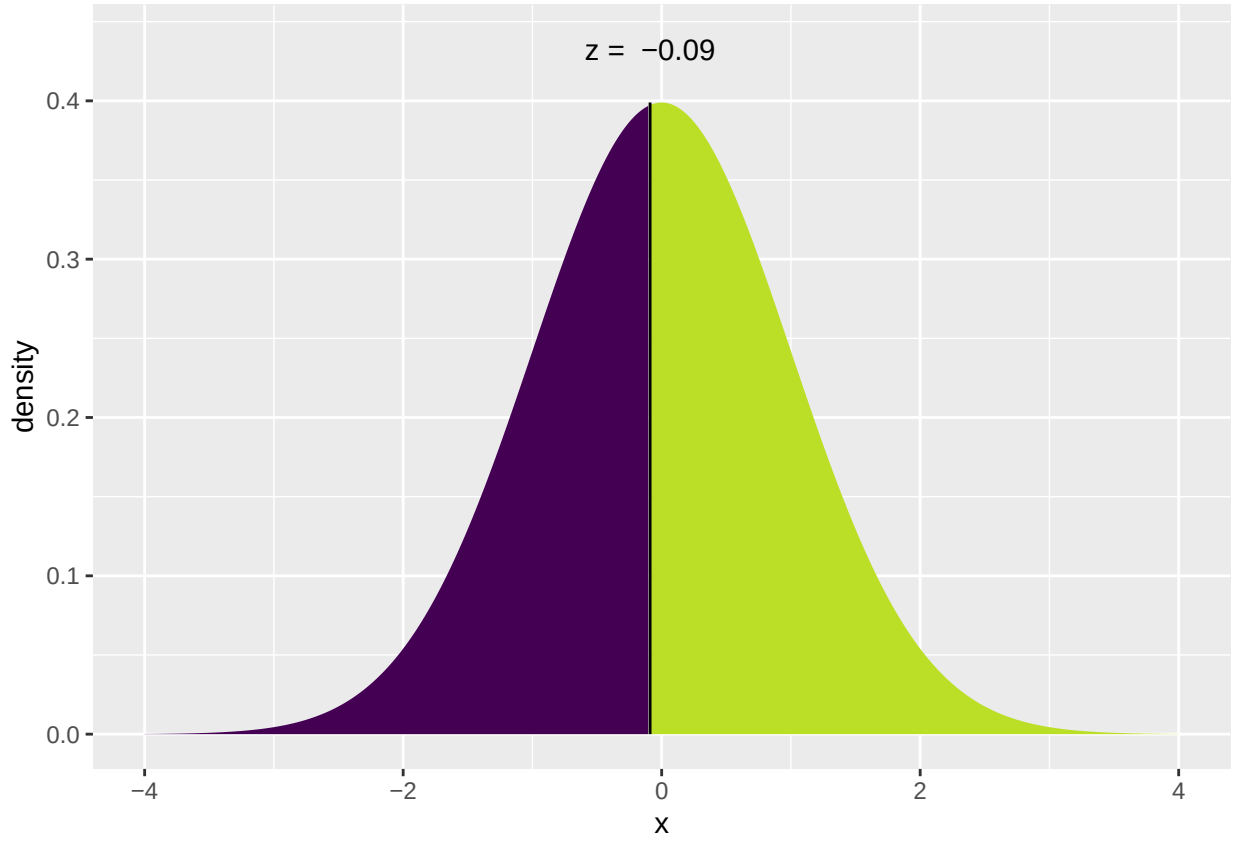
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 55/80
```

```r
US_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                      16
## 2 TRUE                       34
```

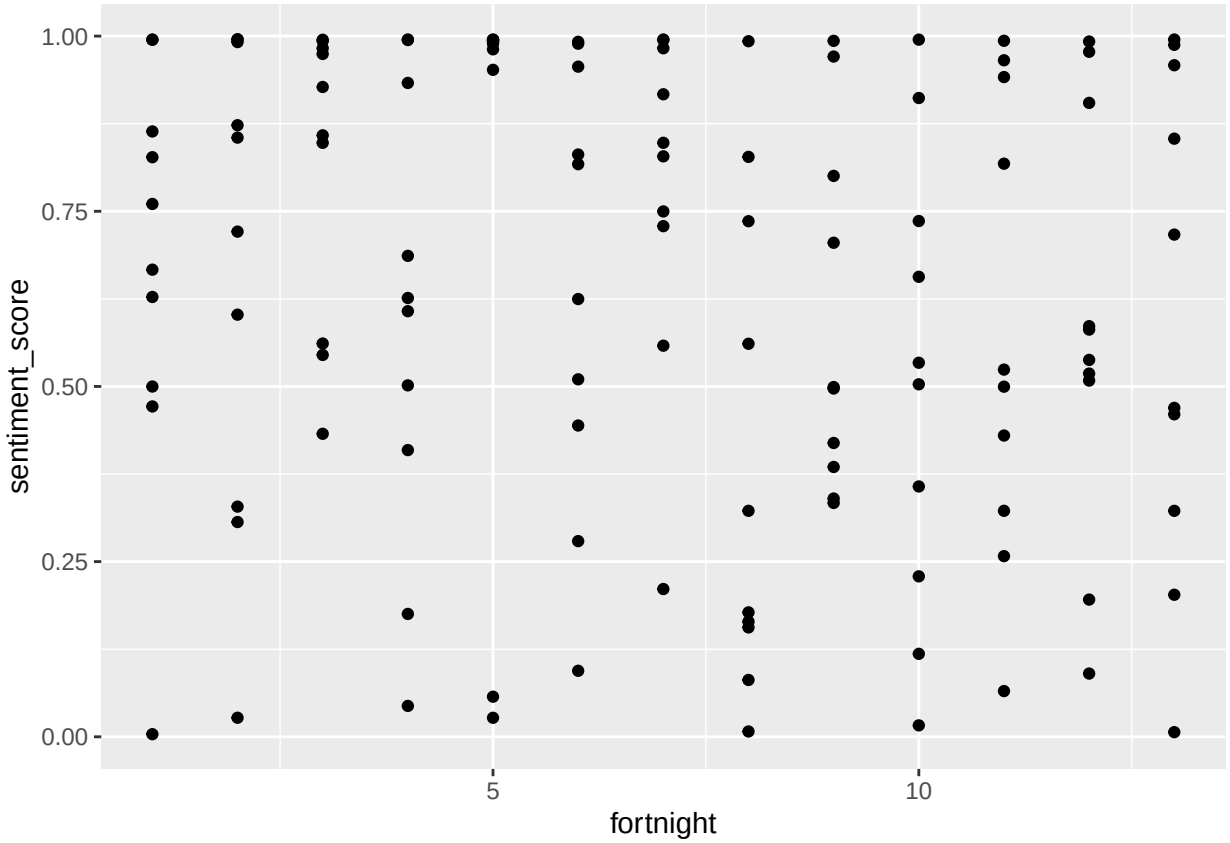
```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 34/50

p_hat = (55+34)/(80+50)

sd <- sqrt((((p_hat)*(1-p_hat))/80)+(((p_hat)*(1-p_hat))/50))
z_score <- ((p_hat2-p_hat1)-0)/sd

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

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.08953) = P(Z <= -0.08953) = 0.4643
##  P(X >  -0.08953) = P(Z >  -0.08953) = 0.5357
##
```

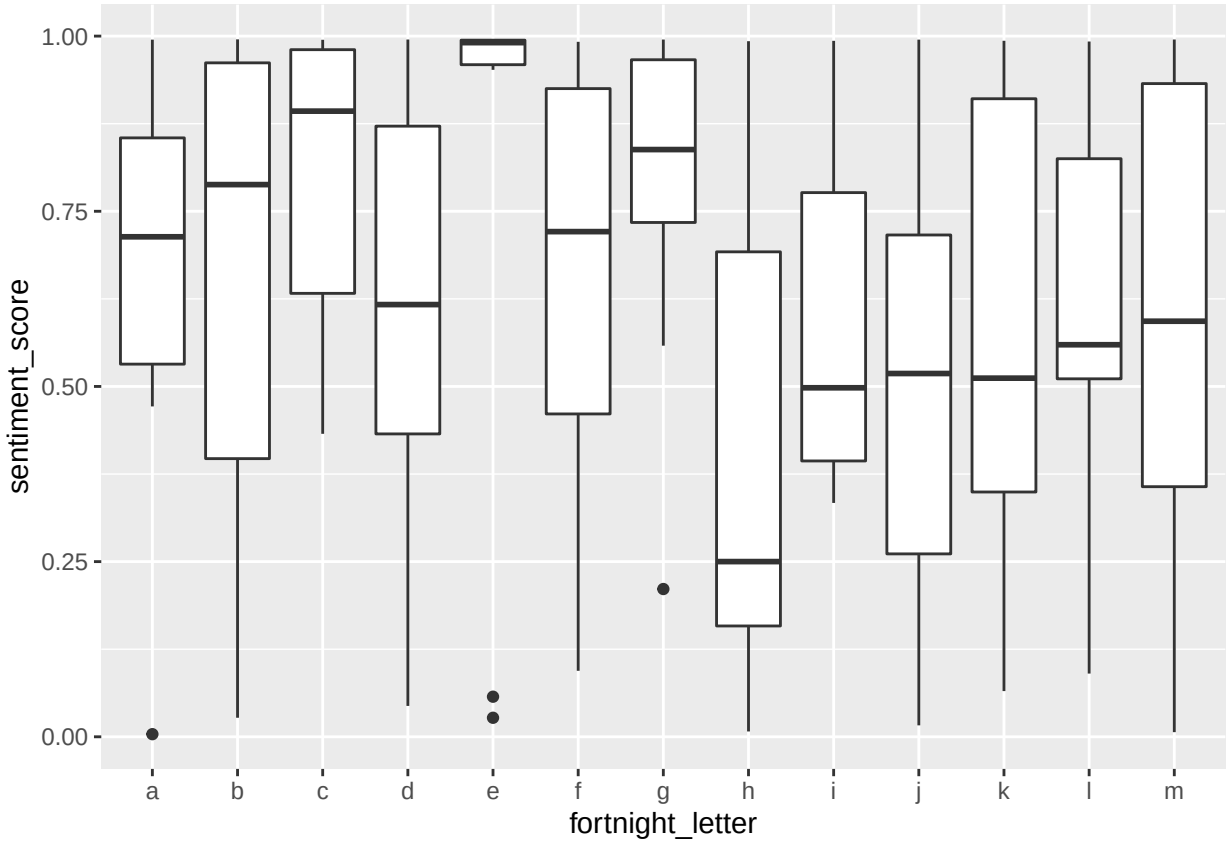



```
## [1] 0.9286595
```

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



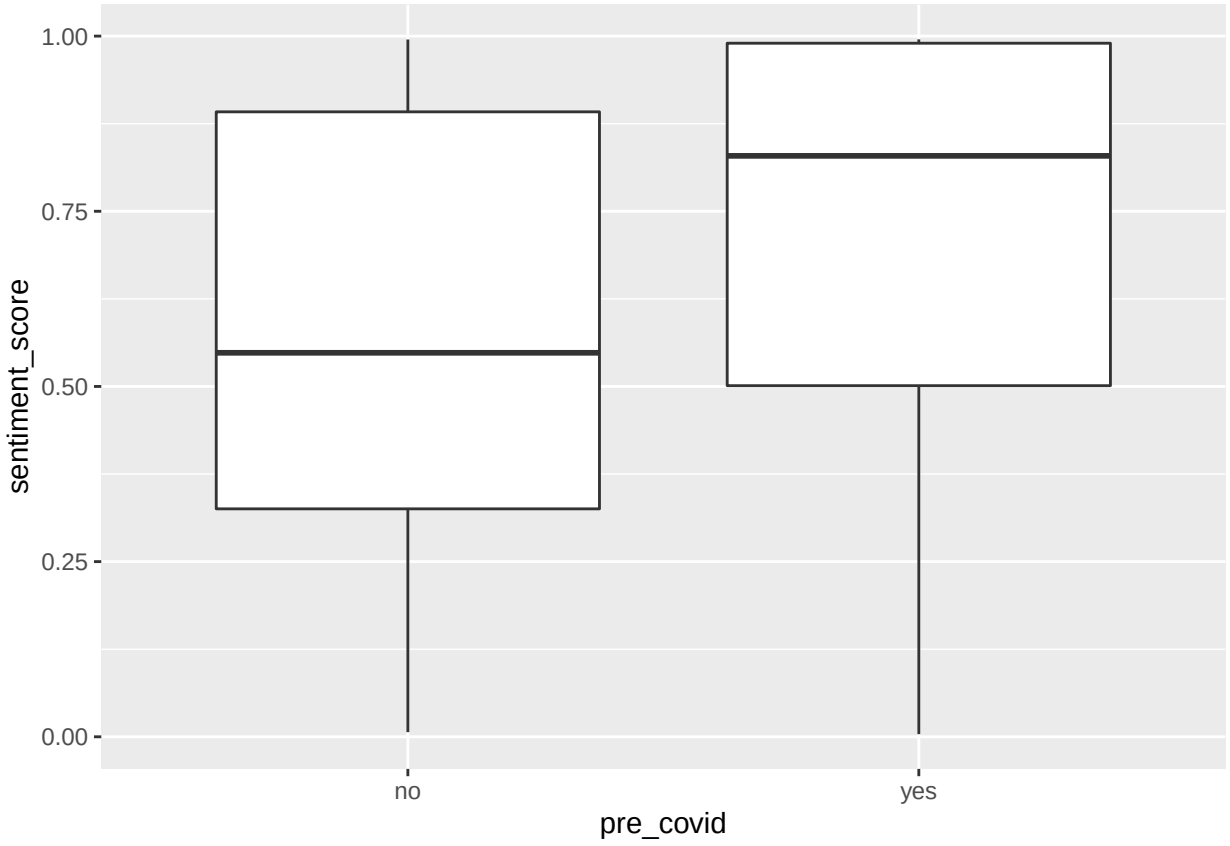

```
ggplot(US_analysis_science) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.671
##  2         2                   0.670
##  3         3                   0.812
##  4         4                   0.597
##  5         5                   0.798
##  6         6                   0.654
##  7         7                   0.781
##  8         8                   0.403
##  9         9                   0.594
## 10        10                   0.506
## 11        11                   0.582
## 12        12                   0.589
## 13        13                   0.597
```

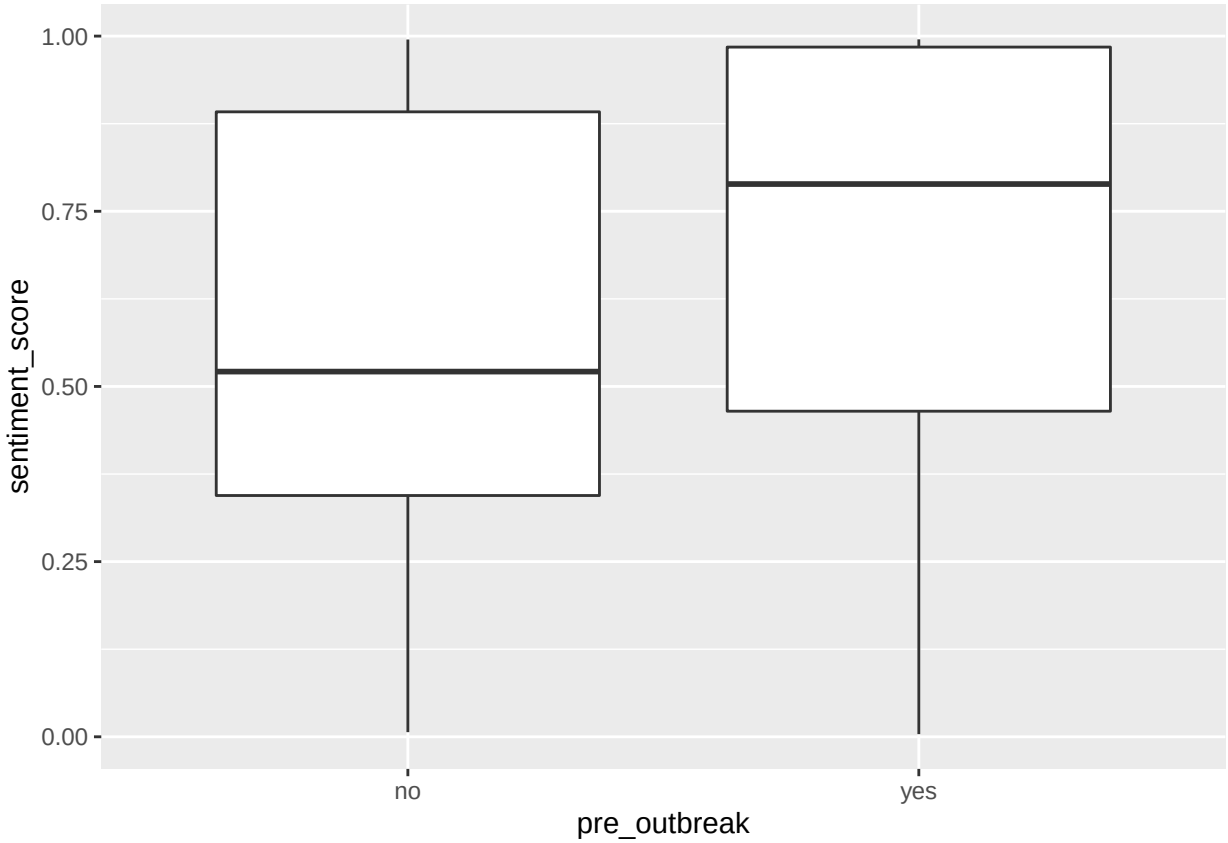
```
ggplot(US_analysis_science) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.579
## 2 yes                         0.700
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.574
## 2 yes                            0.673
```

```
#precovid scitech
count(US_analysis_science, 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

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

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

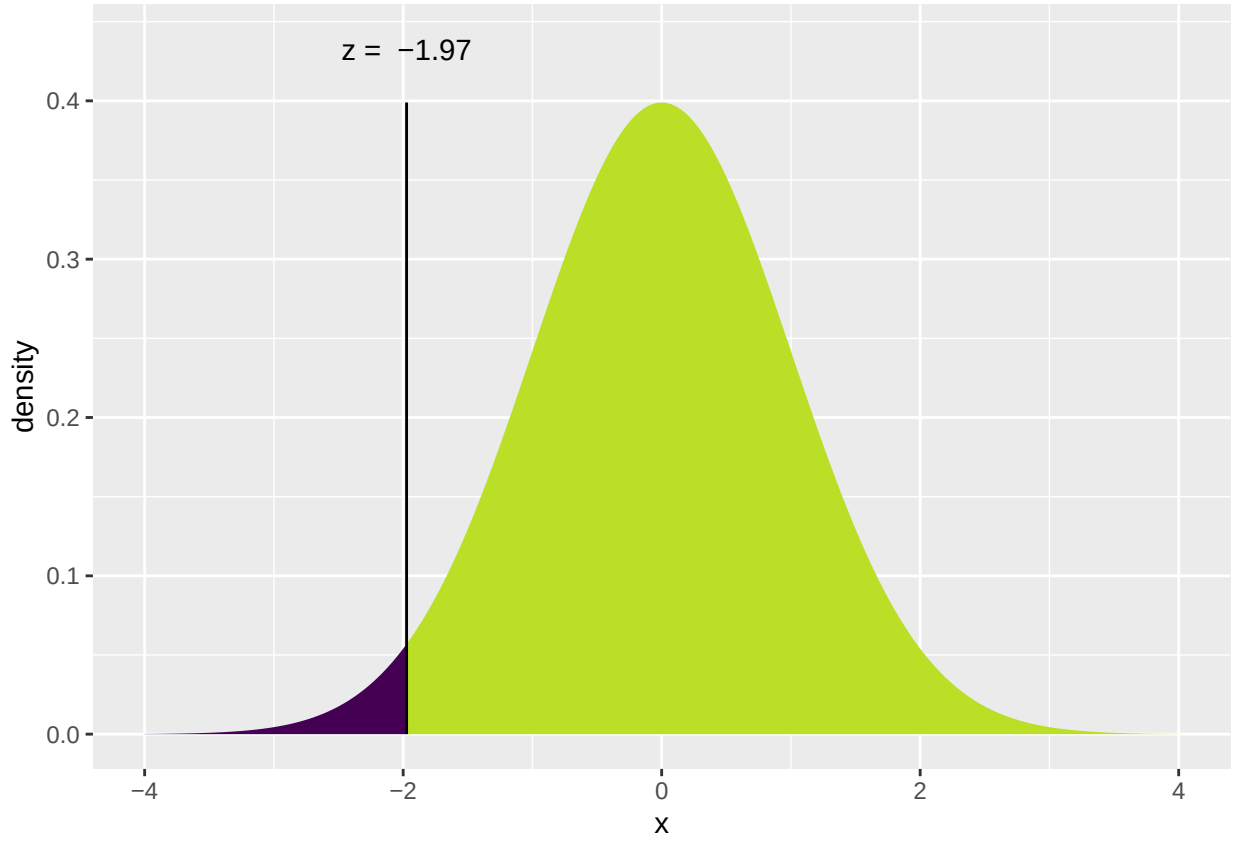
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      15
## 2 TRUE                       45
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 45/60

US_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 = (45+41)/(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* (xpnorm(z_score, 0, 1))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.973) = P(Z <= -1.973) = 0.02422
##  P(X >  -1.973) = P(Z >  -1.973) = 0.9758
##
```




```
## [1] 0.0484472
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      50
## 2 TRUE                       80
```

```
num_preoutbreak = 80
num_postoutbreak = 50
num = 130

US_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                      22
## 2 TRUE                       58
```

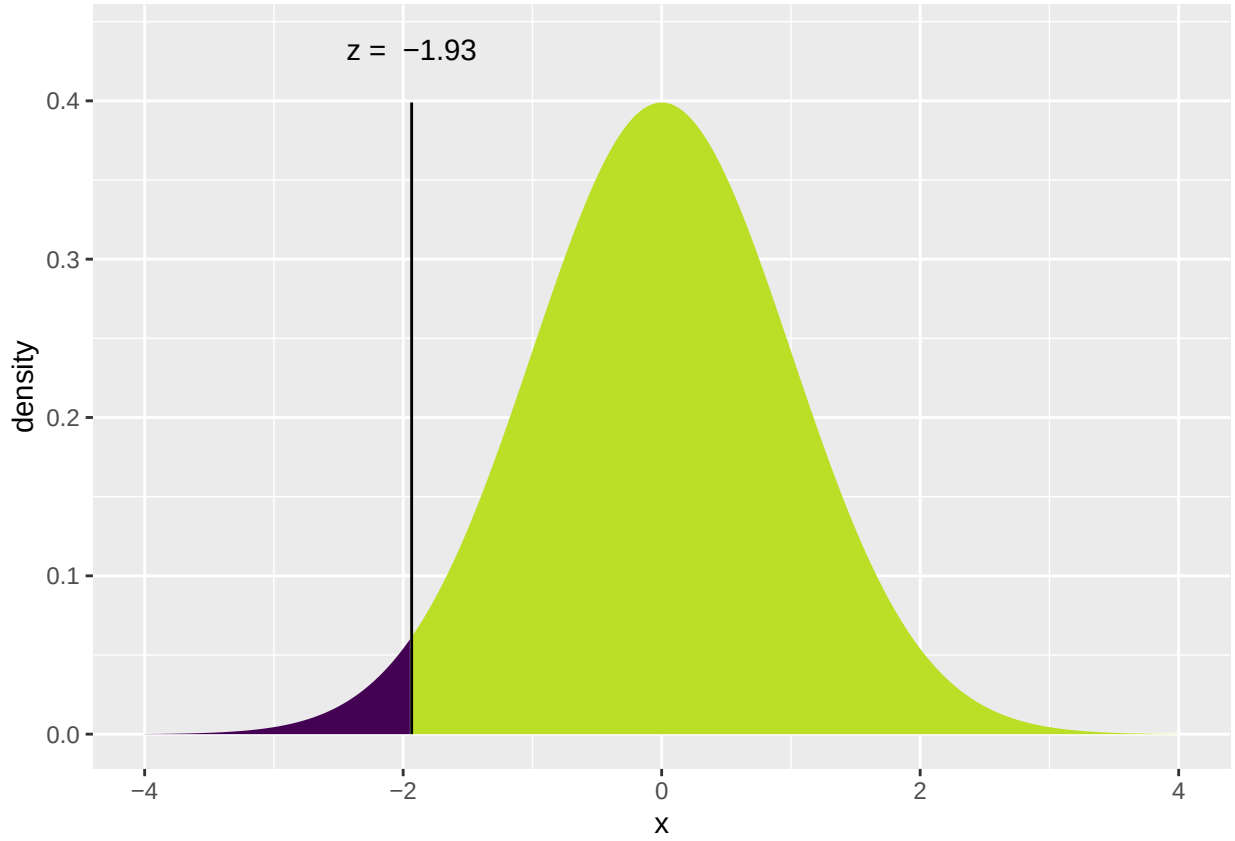
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 58/80
```

```r
US_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                      22
## 2 TRUE                       28
```

```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 28/50

p_hat = (58+28)/(80+50)

sd <- sqrt((((p_hat)*(1-p_hat))/80)+(((p_hat)*(1-p_hat))/50))
z_score <- ((p_hat2-p_hat1)-0)/sd

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

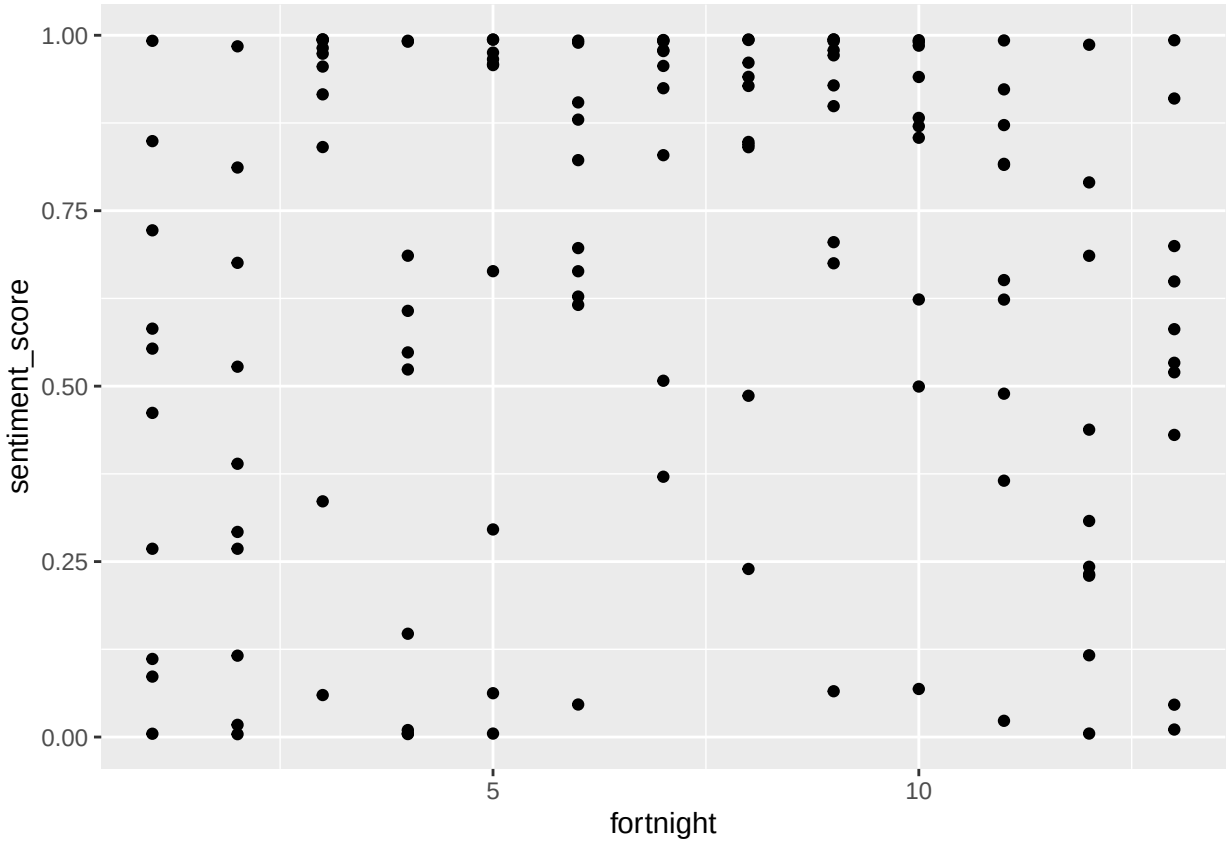
```
##
## If X ~ N(0, 1), then
##  P(X <= -1.934) = P(Z <= -1.934) = 0.02654
##  P(X >  -1.934) = P(Z >  -1.934) = 0.9735
##
```



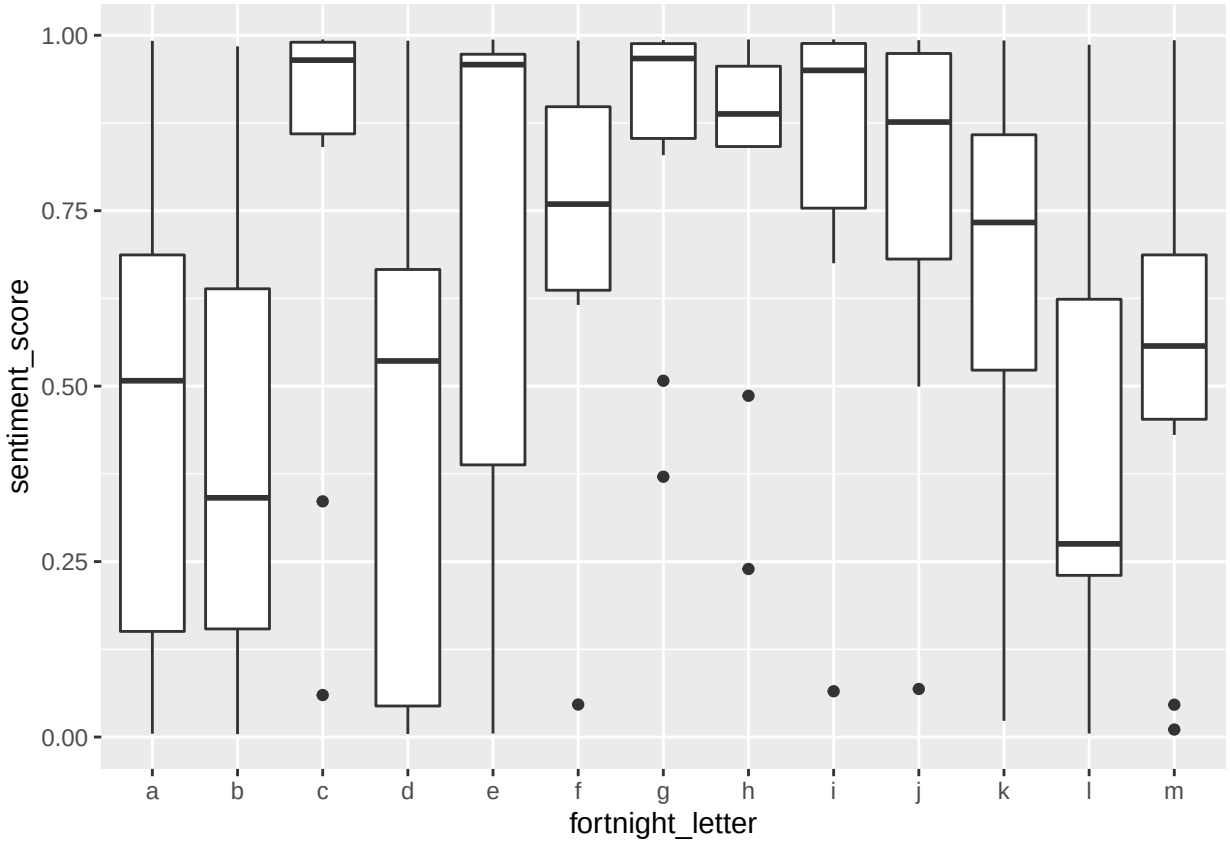

```
## [1] 0.0530838
```

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

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



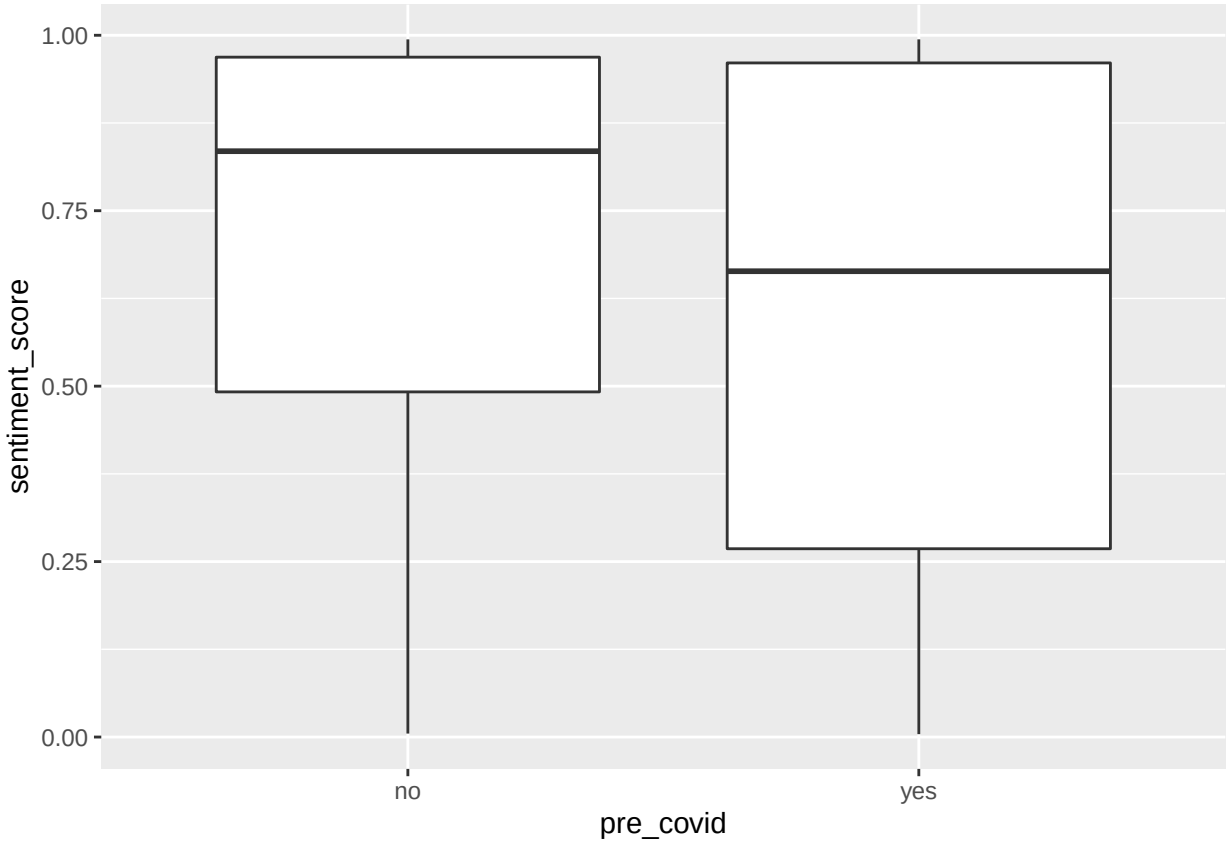

```
ggplot(US_analysis_all) +
  geom_boxplot(aes(x = fortnight_letter, y = sentiment_score))
```




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

```
## # A tibble: 13 x 2
##    fortnight `mean(sentiment_score)`
##        <dbl>                   <dbl>
##  1         1                   0.463
##  2         2                   0.409
##  3         3                   0.804
##  4         4                   0.451
##  5         5                   0.687
##  6         6                   0.724
##  7         7                   0.852
##  8         8                   0.808
##  9         9                   0.820
## 10        10                   0.771
## 11        11                   0.657
## 12        12                   0.403
## 13        13                   0.537
```

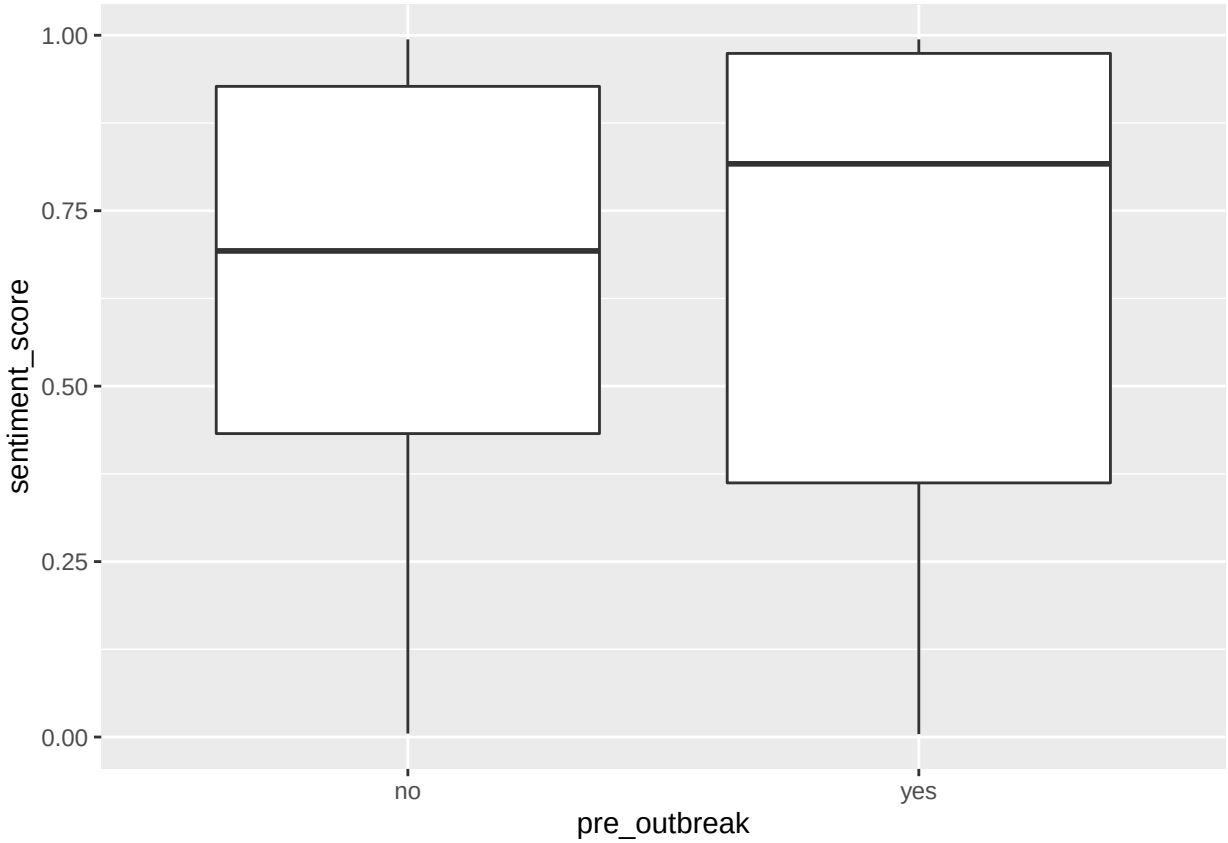
```
ggplot(US_analysis_all) +
  geom_boxplot(aes(x = pre_covid, y = sentiment_score))
```




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

```
## # A tibble: 2 x 2
##   pre_covid `mean(sentiment_score)`
##   <chr>                       <dbl>
## 1 no                          0.693
## 2 yes                         0.590
```

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




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

```
## # A tibble: 2 x 2
##   pre_outbreak `mean(sentiment_score)`
##   <chr>                          <dbl>
## 1 no                             0.638
## 2 yes                            0.650
```

```
#precovid all categories
count(US_analysis_all, 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

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

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

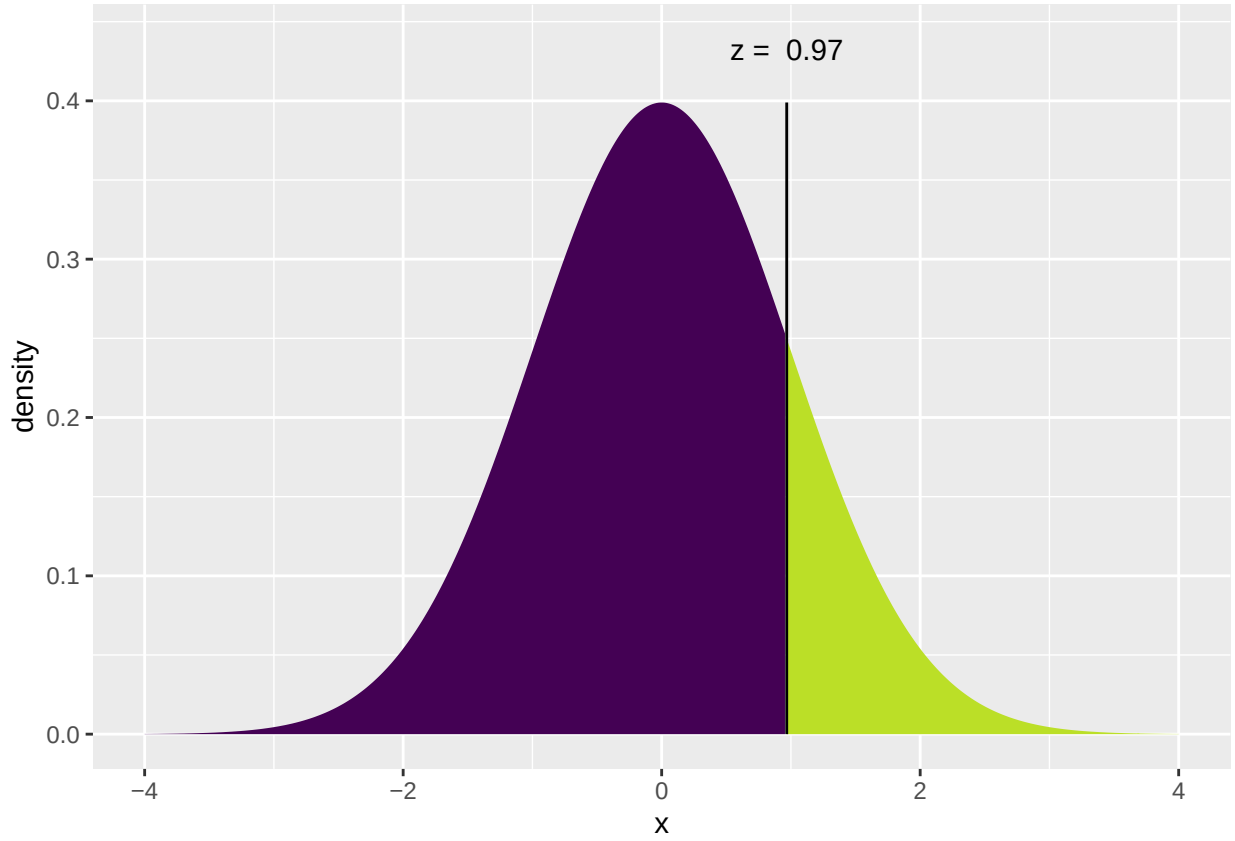
```
##   `sentiment_score > 0.5`     n
##   <lgl>                   <int>
## 1 FALSE                      21
## 2 TRUE                       39
```

```
#proportion of positive sentiment videos precovid from sample
p_hat1 = 39/60

US_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                      19
## 2 TRUE                       51
```

```
#proportion of positive sentiment videos postcovid from sample
p_hat2 =51/70

p_hat = (39+51)/(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 <= 0.9676) = P(Z <= 0.9676) = 0.8334
##  P(X >  0.9676) = P(Z >  0.9676) = 0.1666
##
```




```
## [1] 0.3332287
```

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

```
## # A tibble: 2 x 2
##   `pre_outbreak == "yes"`     n
##   <lgl>                   <int>
## 1 FALSE                      50
## 2 TRUE                       80
```

```
num_preoutbreak = 80
num_postoutbreak = 50
num = 130

US_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                      24
## 2 TRUE                       56
```

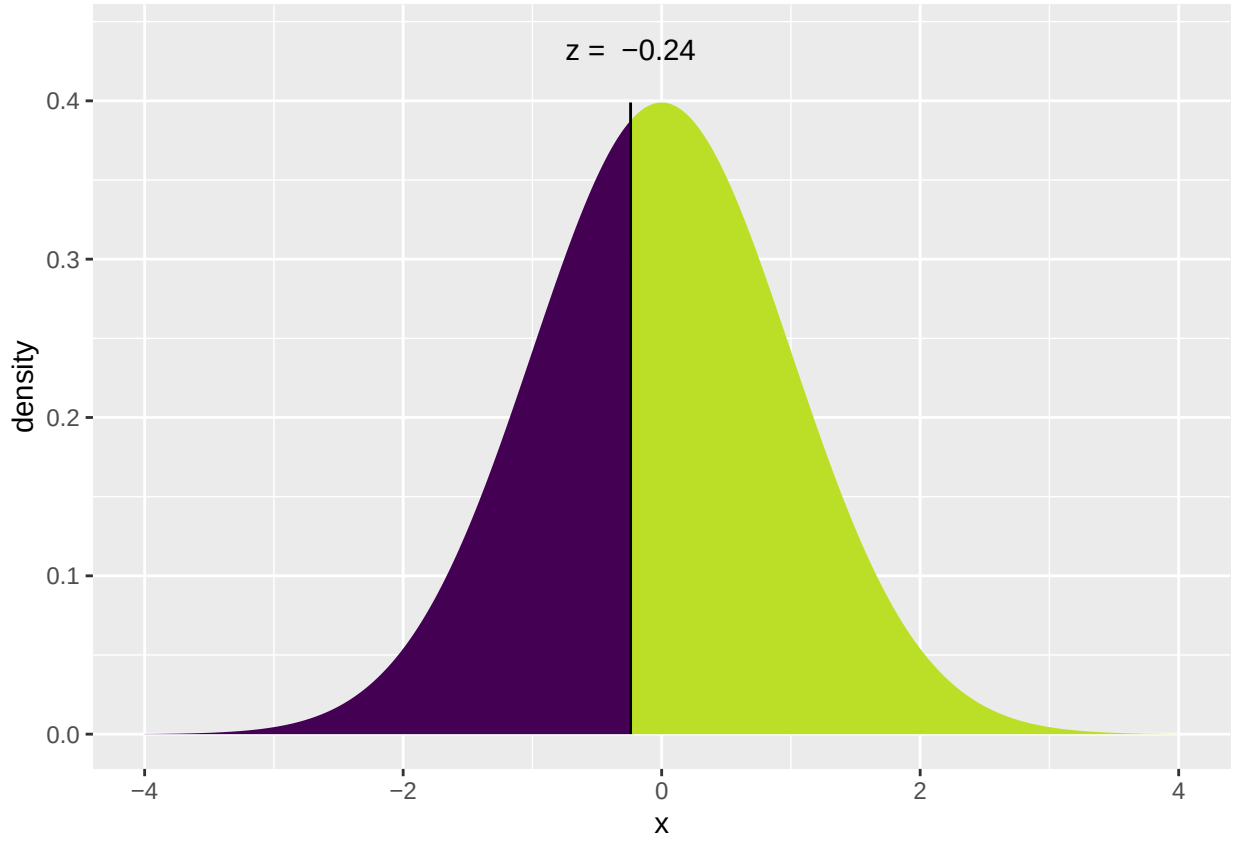
```
#proportion of positive sentiment videos preoutbreak from sample
p_hat1 = 56/80
```

```r
US_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                      16
## 2 TRUE                       34
```

```r
#proportion of positive sentiment videos postoutbreak from sample
p_hat2 = 34/50

p_hat = (56+34)/(80+50)

sd <- sqrt((((p_hat)*(1-p_hat))/80)+(((p_hat)*(1-p_hat))/50))
z_score <- ((p_hat2-p_hat1)-0)/sd

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

```
##
## If X ~ N(0, 1), then
##  P(X <= -0.2404) = P(Z <= -0.2404) = 0.405
##  P(X >  -0.2404) = P(Z >  -0.2404) = 0.595
##
```




```
## [1] 0.8100434
```

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

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.26554, df = 122.07, p-value = 0.791
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1093453  0.1432249
## sample estimates:
##  mean in group no mean in group yes
##         0.5714842         0.5545444
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.20763, df = 106.16, p-value = 0.8359
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.1410260  0.1142872
## sample estimates:
##  mean in group no mean in group yes
##         0.5553141         0.5686835
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.3442, df = 123.54, p-value = 0.1814
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04274451  0.22364743
## sample estimates:
##  mean in group no mean in group yes
##         0.5754868         0.4850353
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 1.3771, df = 97.353, p-value = 0.1716
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04260254  0.23574101
## sample estimates:
##  mean in group no mean in group yes
##         0.5931672         0.4965979
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.23163, df = 127.47, p-value = 0.8172
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08407956  0.10637403
## sample estimates:
##  mean in group no mean in group yes
##         0.7553102         0.7441629
```

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

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

```
## data:  sentiment_score by pre_outbreak
## t = -0.49143, df = 94.86, p-value = 0.6243
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1263419  0.0762046
## sample estimates:
##  mean in group no mean in group yes
##         0.7347385         0.7598071
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.4485, df = 112.72, p-value = 0.1503
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03089568  0.19889013
## sample estimates:
##  mean in group no mean in group yes
##         0.6508300         0.5668328
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.71369, df = 119.51, p-value = 0.4768
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07097912  0.15098488
## sample estimates:
##  mean in group no mean in group yes
##         0.6366792         0.5966764
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.3033, df = 120.9, p-value = 0.195
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04643392  0.22534759
## sample estimates:
##  mean in group no mean in group yes
##         0.5436469         0.4541901
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = 0.66178, df = 105.59, p-value = 0.5096
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09242914  0.18504418
## sample estimates:
##  mean in group no mean in group yes
##         0.5310914         0.4847839
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 0.33917, df = 125.74, p-value = 0.735
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08987314  0.12705021
## sample estimates:
##  mean in group no mean in group yes
##         0.6316795         0.6130910
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.86642, df = 103.57, p-value = 0.3883
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.16017409  0.06277054
## sample estimates:
##  mean in group no mean in group yes
##         0.5931298         0.6418316
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -0.093264, df = 123.83, p-value = 0.9258
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1144924  0.1041882
## sample estimates:
##  mean in group no mean in group yes
##         0.6468118         0.6519638
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.14281, df = 94.887, p-value = 0.8867
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1235517  0.1069695
## sample estimates:
##  mean in group no mean in group yes
##         0.6440874         0.6523785
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = -2.1768, df = 124.91, p-value = 0.03138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.23154610 -0.01101184
## sample estimates:
##  mean in group no mean in group yes
##         0.5789204         0.7001993
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -1.7595, df = 109.71, p-value = 0.08128
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.21146936  0.01256699
## sample estimates:
##  mean in group no mean in group yes
##         0.5736945         0.6731457
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_covid
## t = 1.6947, df = 117.06, p-value = 0.09279
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01734725  0.22313145
```

```
## sample estimates:
##  mean in group no mean in group yes
##         0.6926683         0.5897762
```

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

```
##
##  Welch Two Sample t-test
##
## data:  sentiment_score by pre_outbreak
## t = -0.19637, df = 109.63, p-value = 0.8447
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1332495  0.1092246
## sample estimates:
##  mean in group no mean in group yes
##         0.6377874         0.6497999
```